Fusion or Confusion? Multimodal Complexity Is Not All You Need

Tillmann Rheude[1,3†], **Roland Eils**[1,2,3†], **Benjamin Wild**[1,4†]
[1]Berlin Institute of Health, Charité - Universitätsmedizin Berlin, [2]Intelligent Medicine Institute, Fudan University, [3]Department of Mathematics and Computer Science, Freie Universität Berlin, [4]Institute of Medical Informatics, Charité - Universitätsmedizin Berlin
[†]{benjamin.wild, roland.eils, tillmann.rheude}@bih-charite.de

Abstract

Multimodal learning has become a prominent research area, with the potential of substantial performance gains by combining information across modalities. At the same time, model development has trended toward increasingly complex deep learning architectures, motivated by the assumption that multimodal-specific methods improve performance. We challenge this assumption through a large-scale empirical study by reimplementing 19 high-impact Late Fusion (LF) methods across nine diverse datasets with up to 23 modalities. Under standardized experimental conditions, including hyperparameter tuning, weight initialization, cross-validation, and statistical testing, increased multimodal complexity often yields *confusion rather than effective fusion* of data modalities. Accordingly, complex multimodal architectures do not reliably outperform unimodal baselines and a **Sim**ple **Ba**seline for **M**ulti**m**odal Learning (SimBaMM). Through a focused case study, we further demonstrate concrete methodological shortcomings even in top-tier multimodal learning publications, underscoring the need for standardized evaluation practices. In summary, we argue for a shift in focus for multimodal learning: away from the pursuit of architectural novelty and toward methodological rigor.[1]

1 Introduction

Multimodal learning leverages information from multiple data modalities to improve predictive performance of deep learning architectures. This is especially important for domains with many modalities such as medical applications [1]. Therefore, multimodal learning can enrich unimodal representations and enable models to capture more complex patterns than any single modality alone. At the same time, the pursuit of state-of-the-art (SoTA) performance in multimodal learning has led to increasingly complex architectures. The field is engaged in an architectural arms race, driven by the implicit assumption that novel fusion mechanisms, gradient-based regularizations, and deeper, more sophisticated networks are the primary drivers of progress. This has resulted in a landscape in which models are often celebrated for their structural novelty, with performance gains on specific datasets presented as evidence of their superiority. However, such gains are difficult to interpret when evaluation protocols differ across papers and when key experimental factors*, e.g.,* hyperparameter tuning, weight initialization, and missing-modality simulation, are not controlled or present at all.

In this paper, we challenge the prevailing narrative that architectural novelty is the primary driver of performance in multimodal learning. We argue that methodological rigor in model selection and evaluation can explain a substantial portion of reported improvements. Accordingly, performance gains which are attributed to complex architectural contributions may not reliably persist in fair comparisons. We hypothesize that multimodal learning beyond image-text pairs currently does not

[1]The code repository is uploaded to GitHub.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

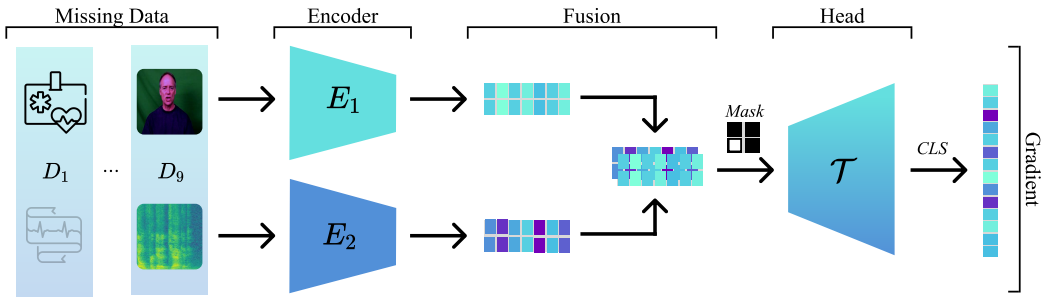


Figure 1: Taxonomy of multimodal learning methods exemplified by two modalities. An LF baseline (SimBaMM) processes a multimodal dataset D with up to M modalities through encoders E_m, combines them via fusion, and produces predictions through a classification head. Categorized into five groups, we evaluate 19 SoTA methods across nine datasets based on their architectural innovations on top of SimBaMM: **Missing Data** handling for incomplete modalities, *e.g.,* a missing ECG, **Encoder**-level modifications, **Fusion** mechanisms for combining representations, **Head**-based cross-modal architectures, and **Gradient**-based optimization techniques. Our work reveals that none of these complex extensions reliably outperforms SimBaMM, suggesting that reported performance gains are often attributable to experimental confounders rather than methodological novelty.

reliably benefit from increasingly complex architectural machinery on top of a simple LF baseline. While multimodal learning can in principle improve downstream performance and should often contain more task-relevant information than any single modality alone, we argue that this potential is currently not captured by more complex multimodal methods.

We perform a large-scale empirical study evaluating 19 multimodal architectures across nine datasets with up to 488 131 samples and 23 modalities. Our goal is to disentangle the contributions of architectural complexity from those of experimental practice. Given this large-scale effort of implementing, tuning and fairly comparing these multimodal methods, our findings reveal that increased architectural complexity does not yield consistent gains over a Simple Baseline for Multimodal Learning (SimBaMM). Rather than asking only in a benchmark-manner which multimodal architecture performs best, we ask under which evaluation conditions such conclusions are scientifically justified. Further, from a machine learning perspective, this finding is striking: In a literature where new multimodal architectures are repeatedly reported as superior to unimodal models, standard baselines, and earlier multimodal methods, these advantages often fail to persist under careful evaluation. If such claims are accepted without rigorous controls, the field risks optimizing for apparent novelty rather than for reliable scientific progress. Taken together, our key contributions are:

- **A large-scale analysis of multimodal architectures under a unified protocol** We reimplement and evaluate 19 methods across nine datasets with up to 23 modalities for controlled comparisons in low- and large-data regimes even beyond the original evaluation settings.
- **A Simple Baseline for Multimodal Learning (SimBaMM)** We show that recent multimodal methods are not able to outperform well-tuned unimodal baselines or a SimBaMM (Figure 1) given fair comparisons and statistical correctness.
- **A case study revealing how experimental confounders inflate performance** We highlight common multimodal evaluation weaknesses, such as information leakage during validation and inadequate hyperparameter tuning, by including a case study of a recent method. This shows how confounders can produce misleading impressions of progress, motivating a shift toward more careful and statistically grounded evaluation in machine learning.

2 Related Work

Multimodal Learning Multimodal learning addresses the use of multiple modalities to perform prediction tasks. We focus on architectural contributions to multimodal learning rather than specific applications and we also exclude vision-language models since they pose a special case of multimodal learning. Architectures in multimodal deep learning can be divided into groups (Figure 1) spanning missing-data handling [56, 74, 77], encoder [19, 30] and fusion mechanisms [8, 39, 46, 55, 83], head

models [71], and gradient regulation methods [57, 75, 79–82]. There is also work on meta-learning [48] and similar extensions such as pre-training or distillation [37] to multimodal learning, which are excluded in our work. Recent work conducted a large-scale study of scaling laws for native multimodal models [65]. This line of work both supports and contextualizes our approach: their results motivate our use of an LF transformer, while our study targets a different gap by investigating contemporary multimodal methods across diverse datasets with statistically rigorous evaluation. Furthermore, our work is situated within the context of multimodal benchmarking, with MultiBench [41] as the most closely related effort. However, we extend and differentiate our work from this in several ways. First, we evaluate generalizability on a curated set of large-scale, contemporary real-world datasets. Second, we assess cross-domain performance directly (rather than splitting in- vs. out-of-domain) to attribute gains to methodological contributions and test their transferability [43]. Third, we standardize key implementation details such as weight initializations. This provides an immediate and actionable baseline and overview for the community, which is notably absent as the MultiBench leaderboard remains unreleased since its publication. Finally, we do not only benchmark methods but ask under which conditions supposed performance gains are scientifically justified.

Multimodal Learning with Missing Modalities Prior work on robustly handling missing modalities is broadly classified into two categories: strategy design (which includes architecture-focused designs and model combinations) and data processing (which involves representation learning and modality imputation) [84]. If missing modalities are not imputed directly at the data level, prior knowledge about missing modalities is used to engineer modules for handling missing modalities. This can*, e.g.,* be done with latent space reconstructions [77] or projections [56], and distribution alignments [74]. On the other hand, Transformers [72] are able to handle missing modalities since they can handle arbitrary sequence lengths. Even if prior work reports performance drops with missing modalities and finds that standard Transformers degrade in this setting [49], balanced training with missing modalities can be sufficient for downstream performance [54]. Yet, a large-scale, domain-independent study of such a Simple Baseline (SimBa) is still lacking.

Reproducibility of Machine Learning Methods Machine learning suffers from a reproducibility crisis [31, 63]. There is prior work focusing on reproducibility in healthcare [53] and guidelines for reproducibility in life sciences [27]. More generally, venues for high-impact machine learning publications are aware of this crisis*, i.e.,* by offering solutions such as reproducibility challenges and reports [58]. Nevertheless, we consider not only reproducibility, but also unsound comparisons*, e.g.,* missing hyperparameter optimizations similar to recent learning rate specific work [34].

Simple Baselines (SimBas) There is a variety of SimBas for diverse tasks. For example, for computer vision, there are SimBas for video restoration [35], tracking [42], image classification [11, 45], video text spotting [24], hand mesh reconstruction [94], graph generation [64], video matting [36], spoken-to-sign-language translation [95], and point cloud classification [21]. More general approaches tackle*, e.g.,* active learning [32], Bayesian uncertainty [51], transfer learning [13], continual learning [6, 59, 60, 92], the traveling salesman problem [85], time series [12], knowledge tracing [47], and instruction fine-tuning [93]. In the context of multimodal learning, there is prior work related to multimodal image segmentation [4], multimodal medical reasoning [29], multimodal embeddings [40], multimodal representation learning [52], multimodal Vision-Language Models (VLMs) [87], multimodal learning with Mixture of Experts (MoE) models [38], and multimodal domain generalization [17]. Prior work hints that there is no benefit of multimodal extensions*, e.g.,* fusion models [68], more complex methods [10], or pre-training in genomics [73]. In summary, such an analysis for multimodal learning across all architectural components is missing.

3 Method

Consider a dataset $D = \{(\{x_{i,m}\}_{m \in \mathcal{M}_i}, y_i)\}_{i=1}^{N}$ consisting of N samples. Each sample i is a (sub)set of M possible modalities and a corresponding target label y_i. The set of available modality indices is $\mathcal{M}_i \subseteq \{1, \ldots, M\}$. For an available modality $m \in \mathcal{M}_i$, the input is $x_{i,m}$.

3.1 Simple Baseline for Multimodal Learning (SimBaMM)

The SimBaMM architecture processes each sample through a multi-stage LF pipeline (Figure 1).

Table 1: Overview of datasets and the respective encoders used in our evaluation. Details in Section F.

Dataset	Size	Modalities	Raw Input	Encoders
Crema-D	7,442	2 (Video, Audio)	✓	2× ResNet
MIMIC HAIM	45,050	2 (Laboratory, X-ray)	✓	MLP, ViT
INSPECT	22,449	2 (CT, EHR)	✗	MLPs
MIMIC Symile	10,345	3 (Laboratory, X-ray, ECG)	✓	MLP, ViT, Transformer
CMU-MOSI	2,199	3 (Video, Audio, Language)	✗	2× Transformer, BERT
CMU-MOSEI	22,856	3 (Video, Audio, Language)	✗	2× Transformer, BERT
CH-SIMS	2,281	3 (Video, Audio, Language)	✓	ViViT, Wav2Vec2, BERT
CH-SIMS 2	4,403	3 (Video, Audio, Language)	✓	ViViT, Wav2Vec2, BERT
UKB	488,131	23 (*e.g.*, Metabolomics, EHR, PRS, Proteomics, Biochemistry)	✓ ✗	MLPs

Encoders For each available modality $m \in \mathcal{M}_i$, the input $x_{i,m}$ is passed through a modality-specific encoder E_m to obtain an embedding $e_{i,m} = E_m(x_{i,m})$. Each $e_{i,m}$ is projected into a shared hidden space of dimension d using a modality-specific linear layer P_m: $\hat{e}_{i,m} = P_m(e_{i,m}) \in \mathbb{R}^d$.

Fusion We form an input sequence S_i from the projected embeddings. To balance performance and fine-grained modeling, we study two variants: (i) **SimBaMM**, which uses a configurable number of tokens per modality (determined by the encoder output and a pooling/tokenization choice), and (ii) **SimBaMM$^{\text{CLS}}$**, which represents each modality by a single token*, e.g.,* a `[CLS]` embedding, yielding a fixed multimodal sequence length of M with padding.

Head The head module is a Transformer $\mathcal{T}$ operating over S_i. To handle missing modalities, we construct an attention mask $M_{\text{attn}}(\mathcal{M}_i)$ such that attention is only computed among tokens originating from available modalities. This enables a controlled comparison between mask-only handling and the additional mechanisms used in prior work. The fused representation is $H_i = \mathcal{T}(S_i, M_{\text{attn}}(\mathcal{M}_i))$. Afterwards, a projection $g(\cdot)$ maps the fused representation H_i to a task-specific output $\hat{y}_i = g(H_i)$. We optimize a task loss $\mathcal{L}(y_i, \hat{y}_i)$*, i.e.,* (binary) cross-entropy.

3.2 Comparison Taxonomy, Method Selection & Reimplementations

We group multimodal methods by the type of innovation they introduce beyond LF. Figure 1 summarizes SimBaMM and the comparison groups used throughout the experiments. We benchmark SimBaMM against a representative set of recent and influential multimodal methods to test whether architectural and objective-level innovations provide consistent gains under a rigorous evaluation protocol. Methods were selected to cover the comparison taxonomy in Figure 1 based on: (a) recency (primarily methods from the last few years), (b) publication in top-tier peer-reviewed machine learning venues, and (c) explicit architectural or objective-based modifications beyond a standard LF baseline. We prioritize methods with sufficiently detailed specifications and/or official reference implementations to enable faithful re-implementations. However, personal selection bias is difficult to eliminate entirely. We made every effort to choose representative methods without regard to expected outcomes. Encoders are dataset-specific but held fixed across all methods (Table 1). Details in Section B.

Generalization to heterogeneous tasks Recent publications for multimodal methods typically assume a fixed number of modalities, specific modality types, and a single downstream task. We adapt all methods to (i) an arbitrary number of modalities, (ii) heterogeneous modality types, and (iii) different tasks and loss functions while preserving the original method's defining components. If a method introduces auxiliary losses, we include them; if auxiliary-loss weighting is not specified, we introduce scalar weights and tune them under the same protocol as other hyperparameters.

Optimization and initialization To reduce the degrees of freedom and improve comparability, we adopt a common optimizer choice whenever compatible with the method. By default, we use the ScheduleFree [15] variant of AdamW and SGD only when required by the method. Weight initializations are standardized across methods to ensure stable convergence across all methods. Details in Section G.

3.3 Experimental Protocols

We argue that multimodal methods can only be compared meaningfully under controlled experimental conditions. Therefore, we standardize the key components of training and evaluation across datasets and methods. The resulting protocols aim to reduce experimental confounding and make differences between methods scientifically interpretable. While most evaluated methods report unimodal baselines, other evaluation protocols are applied inconsistently (Table 2). Moreover, unimodal baselines are not tuned similar to the proposed multimodal methods.

Table 2: Overview of evaluation protocols used by recent multimodal methods. Values are the percentage of surveyed multimodal papers using each protocol.

Evaluation Protocol	% of Methods ↑
Hyperparameter Tuning	0
Statistical Significance	0
Cross-Validation	11
Error Bars	21
Unimodal Baselines	82

Missing-modality protocol We simulate missing modalities at the dataset-split level (rather than per batch) to ensure consistent and reproducible training conditions across epochs [61]. Concretely, we pre-generate a binary missingness mask for all samples in a split. This avoids pre-training or leakage on all available data modalities (Section 4.4). Missing rates are defined per number of missing modalities: for a trimodal dataset, a rate of 15% means 15% of samples miss exactly one and another 15% miss exactly two modalities.

Hyperparameter tuning protocol We first conduct a Bayesian hyperparameter search for SimBaMM over a fixed set of parameters and an extensive search space to determine robust base configurations, *e.g.,* encoder/head dimensions. These base settings are then used as a starting point for tuning the remaining methods, focusing on method-specific components and optimizer parameters under a controlled tuning budget (Section J). As a counterexample, we tune the full hyperparameter space of one method independently of SimBaMM and find that this exhaustive per-method tuning does not materially change the results or conclusions (Section G).

Bayesian statistical comparison We employ Bayesian hierarchical analysis following Benavoli et al. [5], Corani et al. [14] to compare methods across datasets (Section H). This allows us to account for correlations between cross-validation folds and estimate the posterior probability of three outcomes: (1) method A outperforms B by $> 1\%$, (2) method B outperforms A by $> 1\%$, or (3) they are practically equivalent (difference within a $\pm 1\%$ Region of Practical Equivalence (ROPE)).

4 Experiments

4.1 Multimodal Datasets

We evaluate all methods on nine real-world multimodal datasets (Table 1), spanning 2199 to $488\,131$ samples and 2 to 23 modalities. This diverse collection enables a broad and controlled analysis of multimodal architectural complexity across substantially different modality regimes and dataset scales. We intentionally exclude synthetic or otherwise unrealistic datasets in order to ground our conclusions in practically relevant settings. The datasets used in this work are: MIMIC Symile [62], MIMIC Haim [66, 67], INSPECT [28], UK Biobank (UKB) [70], MOSI [89], MOSEI [90], CH-SIMS [88], CH-SIMS v2 [44], and Crema-D [9]. These datasets differ in both their input modalities and data representations, while remaining non-synthetic and representative of real-world tasks. For the UKB, missing modalities are an inherent characteristic. Therefore, we implemented zero imputation for this dataset except for the missing data-based methods. Input modalities are z-score normalized for all datasets. Depending on the modality, we use a Multi-Layer Perceptron (MLP), Vision Transformer (ViT) [18], Transformer [72], BERT [16], Wav2Vec2 [3], ViViT [2], or ResNet [25] as encoders. We report mean and standard deviation (SD) using subject-wise five-fold cross-validation. This eliminates subject overlap between folds, addressing both a frequent methodological weakness in the field (see Section 4.5) and the ability to generalize to new samples. Details in Section F.

4.2 Downstream Task Analysis

We provide a comprehensive empirical downstream comparison of SimBaMM against the 19 re-implemented SoTA methods, including unimodal performances in Figure 2 and Section D.

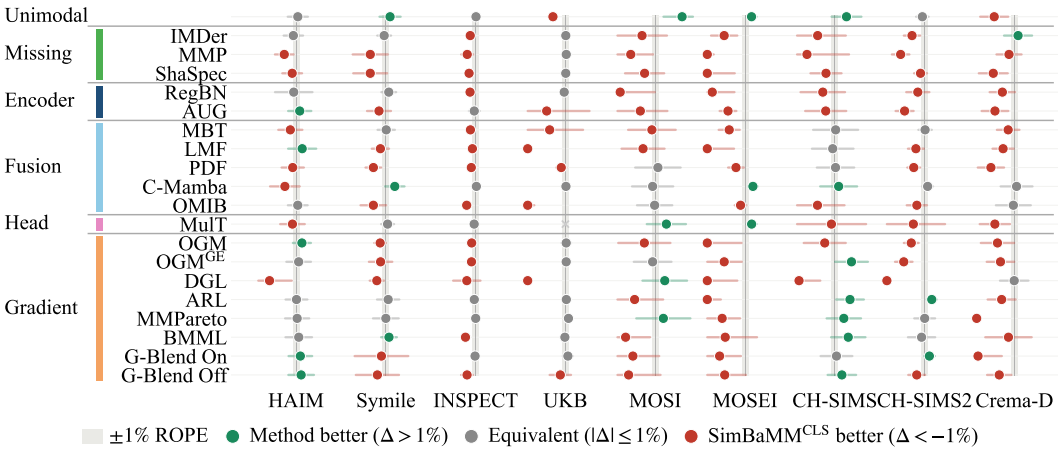


Figure 2: Aggregated forest plot comparing each method against SimBaMM-CLS across all datasets. Points show the mean performance difference relative to SimBaMM-CLS, horizontal lines indicate uncertainty, and the shaded region marks the dataset-specific $\pm 1\%$ ROPE, rather than the generalized ROPE used in the other tables. Unimodal represents the best-performing modality. No method shows a consistent aggregate advantage over SimBaMM-CLS across datasets. Details in Section D.

The forest plot demonstrates that no method is adding significant downstream gains given a dataset-specific $\pm 1\%$ ROPE. Further, there are even outlier methods which do not generalize to datasets with many modalities ($M \gg 2$). This is exemplified with MulT and LMF which become unstable in this setting. MulT scales quadratically, requiring $M(M-1)$ cross-modal transformers (506 for M=23), which is memory-prohibitive. LMF's multiplicative fusion leads to vanishing gradients and activation collapse (AUROC $\approx$ random), suggesting tensor-product fusion is ill-suited for $M \gg 2$. For a statistical, dataset-wide summarizing comparison of all methods, we aim to answer two critical questions (Section 3.3 and Details in Section H):

Does multimodal complexity yield performance gains? Our analysis suggests limited evidence for practically meaningful performance gains from multimodal complexity. No complex method achieves a high probability of outperforming SimBaMM (Table 3). The single highest probability of a complex method winning is $\approx 30\%$ (C-Mamba). Conversely, for methods like AUG, DGL, and OMIB, SimBaMM$^{\text{CLS}}$ is almost certainly superior ($P(\text{Base} > \text{Method}) \geq 95\%$). For the strongest competitors (C-Mamba, MMPareto), the data supports practical equivalence ($P_{\text{ROPE}} \approx 70\%$), indicating limited evidence for a consistent $\geq 1\%$ improvement over the baseline.

Table 3: Comparison with SimBaMM$^{\text{CLS}}$ (*base* in this table). No method achieves a high probability of meaningfully outperforming SimBaMM$^{\text{CLS}}$. Details in Section H.

Method	P(base >other)	P(≈)	P(other >base)	$\mathbb{E}[\delta_0]$	95% CI
Best Unimodal	0.53	0.06	0.41	0.002	[-0.031,0.031]
C-Mamba [39]	0.03	0.67	0.30	-0.006	[-0.021,0.006]
MMPareto [81]	0.13	0.67	0.19	-0.002	[-0.019,0.022]
ARL [80]	0.30	0.47	0.23	0.002	[-0.019,0.028]
GBlend-On [75]	0.36	0.48	0.16	0.005	[-0.018,0.045]
BMML [82]	0.75	0.05	0.19	0.013	[-0.021,0.051]
OGM$^{\text{GE}}$ [57]	0.77	0.16	0.06	0.015	[-0.009,0.042]
MulT [71]	0.33	0.41	0.26	0.002	[-0.023,0.031]
SimBaMM	0.52	0.23	0.25	0.007	[-0.038,0.056]
MBT [55]	0.68	0.29	0.03	0.012	[-0.004,0.028]
GBlend-Off [75]	0.90	0.06	0.04	0.021	[-0.005,0.048]
LMF [46]	0.89	0.01	0.10	0.049	[-0.028,0.142]
OMIB [83]	0.95	0.04	0.01	0.022	[0.003,0.043]
OGM [57]	0.85	0.12	0.03	0.017	[-0.003,0.040]
PDF [8]	0.94	0.05	0.01	0.020	[0.005,0.037]
AUG [30]	0.97	0.01	0.02	0.032	[-0.000,0.066]
RegBN [19]	0.75	0.22	0.03	0.014	[-0.004,0.039]
DGL [79]	0.98	0.00	0.02	0.078	[0.006,0.158]

Do multimodal methods improve upon unimodal baselines? Given our empirical results, multimodal methods do not reliably improve upon unimodal baselines (Table 9). This is consistent with prior observations that the best unimodal model can at times outperform the joint multimodal model [7, 22, 75]. However, in contrast to work that explicitly proposes mechanisms to mitigate such optimization or fusion difficulties [75], we do not propose a dedicated solution here, but instead assess empirically whether

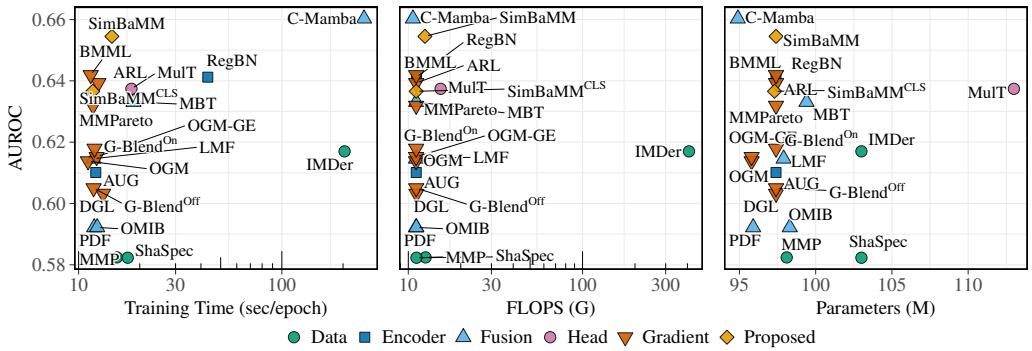


Figure 3: Efficiency analysis for training time (left), FLOPS (middle), and parameters (right) on the MIMIC Symile dataset. Regardless of the metric, SimBaMM and SimBaMM$^{\text{CLS}}$ perform consistently while others*, e.g.,* Coupled Mamba and IMDer are metric-dependent.

the reported multimodal gains hold under a controlled and comparable evaluation setting. While C-Mamba ($\approx 72\%$) and ARL ($\approx 67\%$) likely outperform the best unimodal baseline, many complex architectures*, e.g.,* DGL, OMIB, LMF, and PDF are likely *worse* than simply using the best single modality ($P(\text{Unimodal} > \text{Method}) > 80\%$). SimBaMM$^{\text{CLS}}$ ($\approx 53\%$) rivals the top complex methods in its ability to outperform unimodal baselines, suggesting that architectural complexity does not reliably yield meaningful performance improvements for the multimodal datasets in our study.

4.3 Efficiency Analysis

We analyze the selected methods in terms of Floating Point Operations (FLOPS), number of parameters, and runtime per epoch to demonstrate that many complex models incur high computational costs without yielding improvements in downstream performance (Figure 3). For instance, IMDer [77] requires 203s of runtime and 408 GFLOPS because of the diffusion steps, yet only achieves an AUROC of 0.6170. Similarly, MulT, which has the largest number of parameters (113M), results in a lower AUROC of 0.6374 compared to more parameter-efficient models. While C-Mamba reaches the highest AUROC (0.6603), it comes at the cost of a prohibitive runtime of 253s. In contrast, SimBaMM demonstrates a balance of efficiency and effectiveness. It achieves a high AUROC of 0.6545 while requiring only 14.6s of runtime, 12.5 GFLOPS, and 97.4M parameters. This is $17\times$ faster than C-Mamba for a minimal drop in AUROC. These results underscore that SimBaMM offers strong performance without the substantial computational overhead associated with other methods, positioning it as a practical and effective solution. A parameter-matched analysis was not performed since it is non-trivial to determine how to equitably allocate a fixed parameter budget*, e.g.,* whether to up- or downscale the encoders, the fusion module, or the classifier head. More details in Section E.

4.4 Missing Modality Analysis

Analyzing missing modalities is important because robustness under incomplete input is one of the motivations for multimodal learning. In practice, however, stronger performance of sophisticated methods under missingness cannot be taken for granted: In Figure 4, we compare performance across different missing rates on MIMIC Symile and observe that the competing methods do not reliably outperform SimBaMM. A plausible explanation is that some approaches are trained under assumptions that are not fully aligned with realistic deployment. For example, the public IMDer [77] implementation instructs users to initialize missing-modality experi-

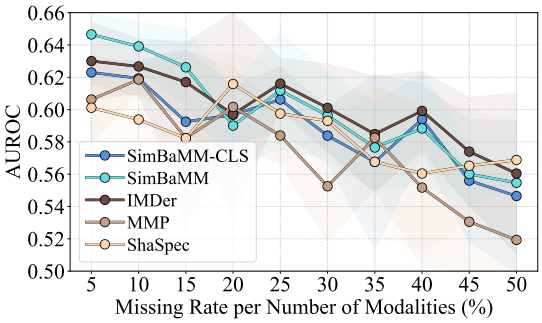


Figure 4: Missing modality analysis for IMDer [77], MMP [56], ShaSpec [74], and SimBaMM. Methods perform similarly across missing rates.

ments from weights pretrained on complete multimodal data, *i.e.,* first using the full modality set and then fine-tuning on the same dataset with missing modalities [78], which may yield a less realistic evaluation protocol for missingness scenarios.

4.5 Case Study

We highlight methodological challenges that can compromise the validity of conclusions to illustrate how common evaluation practices can affect reported results. The Crema-D dataset is widely used in multimodal learning, yet its official repository does not provide standardized cross-validation splits. It states only that subject-independence should be ensured when generating them, leaving this critical decision to individual researchers [76]. We examine AUG[2] [30] as a representative and recent example of how these issues can affect reported results. Based on the original publication, AUG appears to outperform a broad range of multimodal learning methods with an innovative boosting algorithm, supported by ablation studies, bar charts, sensitivity analyses, and t-SNE [50] plots. However, except for tabular results, these validations are reported exclusively on Crema-D. AUG follows the splits from OGM, which uses only training and test sets, resulting in model selection on the test set. Furthermore, the splits are not subject-independent despite the dataset repository's explicit guidance, resulting in subject leakage. Additionally, different optimizers are used across datasets, complicating fair comparisons. When we introduce a standardized dataset split protocol, generalize the implementation of AUG to M modalities, and tune hyperparameters, we observe that AUG continues to boost unimodal video performance. However, it does not improve multimodal performance relative to simpler baselines. Consequently, the relative ordering shifts from AUG outperforming SimBaMM to SimBaMM matching or exceeding AUG (Figure 5). While acknowledging the theoretical contributions of AUG, this demonstrates how common evaluation practices can influence conclusions. Details in Section G.

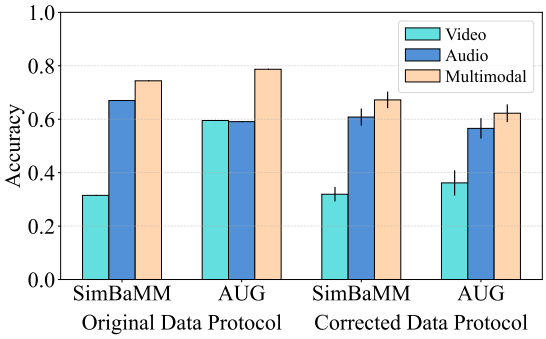


Figure 5: Case study on the CREMA-D dataset. Under a corrected data protocol with subject-independent splits and consistent hyperparameter tuning, the relative performance of methods changes.

5 Discussion

5.1 Limitations

Scope of the work Our conclusions should be interpreted within the scope of the present work. Although we evaluate 19 multimodal methods across nine datasets, our study does not cover all multimodal learning paradigms such as pretraining. Therefore, Early Fusion (EF) methods, large-scale pre-training-based approaches, contrastive objectives, and some domain-specific architectures fall outside the scope of this work. We do not claim that architectural complexity is never useful in multimodal learning, but only that we do not observe reliable practical gains over a strong SimBa in the settings studied. While the evaluated datasets span diverse modality combinations and up to 23 modalities, they do not justify universal conclusions about all multimodal learning problems.

Task settings and biases Our work necessarily involves a trade-off between comparability and method-specific optimization. Some methods were originally proposed for narrower task settings or modality combinations and may therefore be disadvantaged. Similarly, our use of a standardized tuning protocol improves fairness and comparability, but may not be optimal for every architectural family. Starting from infrastructure choices that perform well for SimBaMM may introduce a tuning bias, even though all methods are subsequently tuned with substantial compute budgets. We do not claim that each method performs identically under its original protocol. This would reintroduce multiple confounders. Our conclusions apply to the controlled setting studied here.

[2]Published at NeurIPS 2025 with an oral presentation.

Evaluation protocols Our conclusions depend on concrete evaluation choices, including preprocessing, evaluation protocols, and the use of specific ROPEs in the Bayesian analysis. While we provide sensitivity analyses, alternative choices could affect absolute performance and rankings.

Mechanistic insights Our paper is primarily empirical and diagnostic. It establishes that carefully tuned SimBas are difficult to outperform under controlled evaluation, but it provides only limited mechanistic insight into why some complex methods fail, why certain methods such as Coupled Mamba show positive trends, or which specific design principles should guide future multimodal architectures. We view these questions as important and promising directions for future work.

5.2 Main Drivers of Multimodal Learning

Given our empirical results that complex SoTA multimodal models do not yield reliable performance gains, a critical question arises: If methodological novelty is not the answer, what are the true drivers of multimodal performance? We argue that most problems framed as multimodal are, in fact, general machine learning challenges. Consider the core motivations behind the methods in our study: non-alignment, differential learning dynamics across modalities, dominant modality suppression, missing modality imputation, and distribution heterogeneity. These challenges arise equally in unimodal learning when combining heterogeneous feature groups, or when learning from features with varying predictive strength. The distinction between *modalities* and *feature groups within a modality* is largely semantic. If these problems do not require specific solutions in unimodal contexts, it is unclear why they would require such solutions in multimodal contexts. This suggests that the performance of any multimodal architecture is anchored to the quality of its unimodal representations. For example, in unimodal learning, the paradigm shift from Convolutional Neural Networks (CNNs) [33] to ViTs [18] improved visual feature extraction. A multimodal model using a ViT backbone benefits from this, even before any fusion is applied. Similarly, innovations in attention*, e.g.,* local, global, or windowed attention, are developed in unimodal contexts (*e.g.*, the Global Context ViT [23]) before possibly being adapted to multimodal architectures. The gains from these components are substantial, yet they are not multimodal innovations. The true performance drivers may be the unimodal parts, while the specific choice of a more complex multimodal method may offer only marginal, if any, benefit.

5.3 Reliability Checklist for Multimodal Learning

We argue for a shift towards reliable results instead of added complexity. For comparable and trustworthy future work, we recommend the following pragmatic best practices:

(1) Standardize experimental conditions: Use the same optimizers, weight initialization, data splits, and training protocols across all methods. **(2) Use identical base architectures**: When claiming benefit from a specific contribution, use an identical base architecture for both the proposed method and baselines. **(3) Include baselines**: Include simple and unimodal baselines such as SimBaMM to verify that observed gains stem from the proposed contribution rather than confounding factors. **(4) Tune all methods rigorously**: Perform hyperparameter tuning for both the proposed method and baselines. **(5) Use proper cross-validation and reporting**: Employ cross-validation with subject-independent splits and report mean $\pm$ SD rather than single-run results. **(6) Evaluate across multiple diverse datasets**: Assess generalizability by testing methods across multiple diverse and preferably large datasets.

6 Conclusion & Future Work

We challenged the pursuit of architectural novelty in multimodal learning with a large-scale empirical study reimplementing 19 SoTA models across nine diverse datasets with up to 23 modalities. Our analysis reveals that complex methods were at best practically equivalent to our proposed SimBaMM, with none showing strong evidence of meaningful performance gains. Our findings motivate a shift in emphasis: from architectural novelty alone to rigorous evaluation that verifies generalizable improvements. In many settings we study, careful tuning and strong baselines explain most of the observed gains. We demonstrate that performance is more robustly driven by rigorous hyperparameter tuning, sound methodological evaluation, and strong unimodal encoders, rather than by multimodal architectural novelty. Future work could include a similar analysis of other model families, such as EF and MoE architectures, as well as methods employing other training strategies like pre-training,

contrastive learning, or meta-learning. Another objective could be to develop a unified, all-inclusive model based on SimBaMM.

Acknowledgement We would like to thank Georg v. Arnim for the support with preprocessing the INSPECT dataset. Further, we would like to thank Stefan Hegselmann, Tom Burgert, Leon Sixt and Jake Graving for their valuable feedback and discussions on this paper. The authors acknowledge the Scientific Computing of the IT Division at the Charité - Universitätsmedizin Berlin for providing computational resources that have contributed to the research results reported in this paper. This research has been conducted using the UK Biobank Resource under application number 49966. B.W. and R.E. acknowledge support by the Collaborative Research Center (SFB 1470) funded by the German Research Council (DFG). The authors gratefully acknowledge the computing time made available to them on the high-performance computer "Lise" at the NHR Center NHR@ZIB. This center is jointly supported by the Federal Ministry of Education and Research and the state governments participating in the NHR (`www.nhr-verein.de/unsere-partner`).

References

[1] Julián N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, September 2022. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-022-01981-2.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, page 6816–6826. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00676. URL `https://doi.org/10.1109/ICCV48922.2021.00676`.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html`.

[4] Matteo Bastico, David Ryckelynck, Laurent Corté, Yannick Tillier, and Etienne Decencière. A simple and robust framework for cross-modality medical image segmentation applied to vision transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, page 4130–4140. IEEE, 2023. doi: 10.1109/ICCVW60793.2023.00446. URL `https://doi.org/10.1109/ICCVW60793.2023.00446`.

[5] Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18(77):1–36, 2017.

[6] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and SIMONE CALDERARA. Dark experience for general continual learning: a strong, simple baseline. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, page 15920–15930. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/b704ea2c39778f07c617f6b7ce480e9e-Paper.pdf`.

[7] Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases for visual question answering. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/51d92be1c60d1db1d2e5e7a07da55b26-Paper.pdf`.

[8] Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan

Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, page 5608–5628. PMLR, July 2024. URL https://proceedings.mlr.press/v235/cao24c.html.

[9] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

[10] Isabella Caranzano, Corrado Pancotti, Cesare Rollo, Flavio Sartori, Pietro Liò, Piero Fariselli, and Tiziana Sanavia. Sparsity is all you need: Rethinking biological pathway-informed approaches in deep learning. arXiv:2505.04300 [q-bio], May 2025. URL http://arxiv.org/abs/2505.04300.

[11] Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma. Pcanet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, 24(12): 5017–5032, December 2015. ISSN 1057-7149, 1941-0042. doi: 10.1109/TIP.2015.2475625. arXiv:1404.3606 [cs].

[12] Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. Simpletm: A simple baseline for multivariate time series forecasting. In *Proceedings of the 13th International Conference on Learning Representations*, Singapore, 2025. URL https://openreview.net/pdf?id=oANkBaVci5.

[13] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, page 5110–5120. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00506. URL https://doi.org/10.1109/CVPR52688.2022.00506.

[14] Giorgio Corani, Alessio Benavoli, Janez Demšar, Francesca Mangili, and Marco Zaffalon. Statistical comparison of classifiers through bayesian hierarchical modelling. *Machine Learning*, 106:1817–1837, 2017.

[15] Aaron Defazio, Xingyu Yang, Ahmed Khaled, Konstantin Mishchenko, Harsh Mehta, and Ashok Cutkosky. The road less scheduled. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/136b9a13861308c8948cd308ccd02658-Abstract-Conference.html.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, page 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

[17] Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. Simmmdg: A simple and effective framework for multi-modal domain generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f88bec15cc4cb56b432ee040bb63f94f-Paper-Conference.pdf.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[19] Morteza Ghahremani and Christian Wachinger. Regbn: Batch normalization of multimodal data with regularization. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL https://arxiv.org/pdf/2310.00641.

[20] Varun Godbole, George E. Dahl, Justin Gilmer, Christopher J. Shallue, and Zachary Nado. Deep learning tuning playbook, 2023. URL http://github.com/google-research/tuning_playbook. Version 1.0.

[21] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, page 3809–3820. PMLR, 2021. URL http://proceedings.mlr.press/v139/goyal21a.html.

[22] Yu Gu, Jingjing Fu, Xiaodong Liu, Jeya Maria Jose Valanarasu, Noel CF Codella, Reuben Tan, Qianchu Liu, Ying Jin, Sheng Zhang, Jinyu Wang, Rui Wang, Lei Song, Guanghui Qin, Naoto Usuyama, Cliff Wong, Hao Cheng, HoHin Lee, Praneeth Sanapathi, Sarah Hilado, Tristan Naumann, Javier Alvarez-Valle, Jiang Bian, Mu Wei, Khalil Malik, Lidong Zhou, Jianfeng Gao, Eric Horvitz, Matthew P. Lungren, Doug Burger, Eric Topol, Hoifung Poon, and Paul Vozila. The illusion of readiness in health ai. arXiv:2509.18234 [cs], December 2025. URL http://arxiv.org/abs/2509.18234.

[23] Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Global context vision transformers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, page 12633–12646. PMLR, 2023. URL https://proceedings.mlr.press/v202/hatamizadeh23a.html.

[24] Haibin He, Maoyuan Ye, Jing Zhang, Juhua Liu, Bo Du, and Dacheng Tao. Gomatching: A simple baseline for video text spotting via long and short term matching. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2d66a70c770de7835678f1c1e65fe5e1-Paper-Conference.pdf.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, page 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

[26] Stefan Hegselmann, Georg von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful electronic health record encoders. arXiv:2502.17403 [cs], October 2025. URL http://arxiv.org/abs/2502.17403.

[27] Benjamin J. Heil, Michael M. Hoffman, Florian Markowetz, Su-In Lee, Casey S. Greene, and Stephanie C. Hicks. Reproducibility standards for machine learning in the life sciences. *Nature Methods*, 18(10):1132–1135, October 2021. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-021-01256-7.

[28] Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P. Lungren, Curtis Langlotz, Serena Yeung, Nigam Shah, and Jason Alan Fries. Inspect: A multimodal dataset for pulmonary embolism diagnosis and prognosis. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=3sRR2u72oQ.

[29] Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. Medvlthinker: Simple baselines for multimodal medical reasoning. arXiv:2508.02669 [cs], August 2025. URL http://arxiv.org/abs/2508.02669.

[30] Qing-Yuan Jiang, Longfei Huang, and Yang Yang. Rethinking multimodal learning from the perspective of mitigating classification ability disproportion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Q6IyUpBmrG.

[31] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, September 2023. ISSN 26663899. doi: 10.1016/j.patter.2023.100804.

[32] Andreas Kirsch, Sebastian Farquhar, Parmida Atighehchian, Andrew Jesson, Frédéric Branchaud-Charron, and Yarin Gal. Stochastic batch acquisition: A simple baseline for deep active learning. *Transactions on Machine Learning Research*, 2023, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=vcHwQyNBjW.

[33] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14539.

[34] Yu-Ang Lee, Ching-Yun Ko, Pin-Yu Chen, and Mi-Yen Yeh. Learning rate matters: Vanilla lora may suffice for llm fine-tuning. arXiv:2602.04998 [cs.LG], 2026. URL https://arxiv.org/abs/2602.04998.

[35] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, page 9822–9832. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00947. URL https://doi.org/10.1109/CVPR52729.2023.00947.

[36] Jiachen Li, Marianna Ohanyan, Vidit Goel, Shant Navasardyan, Yunchao Wei, and Humphrey Shi. Videomatt: A simple baseline for accessible real-time video matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, page 2177–2186, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/MobileAI/papers/Li_VideoMatt_A_Simple_Baseline_for_Accessible_Real-Time_Video_Matting_CVPRW_2023_paper.pdf.

[37] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, page 9694–9705, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/505259756244493872b7709a8a01b536-Abstract.html.

[38] Sijie Li, Chen Chen, and Jungong Han. Simmlm: A simple framework for multi-modal learning with missing modality. *CoRR*, abs/2507.19264, 2025. doi: 10.48550/ARXIV.2507.19264. URL https://doi.org/10.48550/arXiv.2507.19264. arXiv: 2507.19264.

[39] Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multimodal fusion with coupled state space model. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, page 59808–59832. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6e09c213ac18d6375704a4f3ea75c4f8-Paper-Conference.pdf.

[40] Paul Pu Liang, Yao Chong Lim, Yao-Hung Hubert Tsai, Ruslan Salakhutdinov, and Louis-Philippe Morency. Strong and simple baselines for multimodal utterance embeddings. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://arxiv.org/pdf/1906.02125.

[41] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://arxiv.org/abs/2107.07502.

[42] Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/file/6a5c23219f401f3efd322579002dbb80-Paper-Conference.pdf.

[43] Zachary C. Lipton and Jacob Steinhardt. Research for practice: troubling trends in machine-learning scholarship. *Communications of the ACM*, 62(6):45–53, May 2019. ISSN 0001-0782, 1557-7317. doi: 10.1145/3316774.

[44] Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module. In Raj Tumuluri, Nicu Sebe, Gopal Pingali, Dinesh Babu Jayagopi, Abhinav Dhall, Richa Singh, Lisa Anthony, and Albert Ali Salah, editors, *International Conference on Multimodal Interaction, ICMI 2022, Bengaluru, India, November 7-11, 2022*, page 247–258. ACM, 2022. doi: 10.1145/3536221.3556630.

[45] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, page 11966–11976. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01167. URL https://doi.org/10.1109/CVPR52688.2022.01167.

[46] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, page 2247–2256. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1209. URL https://aclanthology.org/P18-1209/.

[47] Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. simplekt: A simple but tough-to-beat baseline for knowledge tracing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=9HiGqC9C-KA.

[48] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, page 2302–2310. AAAI Press, 2021. doi: 10.1609/AAAI.V35I3.16330. URL https://doi.org/10.1609/aaai.v35i3.16330.

[49] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, page 18156–18165. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01764. URL https://doi.org/10.1109/CVPR52688.2022.01764.

[50] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[51] Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, page 13132–13143, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/118921efba23fc329e6560b27861f0c2-Abstract.html.

[52] Andrei-Marian Manolache, Dragos-Constantin Tantaru, and Mathias Niepert. Molmix: A simple yet effective baseline for multimodal molecular representation learning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, Machine Learning in Structural Biology Workshop, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL https://arxiv.org/abs/2410.07981.

[53] Matthew McDermott. The (lack of?) science of machine learning for healthcare. In Stefan Hegselmann, Helen Zhou, Elizabeth Healey, Trenton Chang, Caleb Ellington, Vishwali Mhasawade, Sana Tonekaboni, Peniel Argaw, and Haoran Zhang, editors, *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, page 19–29. PMLR, December 2025. URL https://proceedings.mlr.press/v259/mcdermott25a.html.

[54] Marius Memmel, Roman Bachmann, and Amir Zamir. Modality-invariant visual odometry for embodied vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 21549–21559, 2023.

[55] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, page 14200–14213, 2021. URL https://arxiv.org/pdf/2107.00135.

[56] Niki Nezakati, Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and M. Salman Asif. Mmp: Towards robust multi-modal learning with masked modality projection. arXiv:2410.03010 [cs], October 2024. URL http://arxiv.org/abs/2410.03010.

[57] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, page 8228–8237. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00806. URL https://doi.org/10.1109/CVPR52688.2022.00806.

[58] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research(a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

[59] Ameya Prabhu, Philip H. S. Torr, and Puneet K. Dokania. Gdumb: A simple approach that questions our progress in continual learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, page 524–540. Springer, 2020. doi: 10.1007/978-3-030-58536-5_31. URL https://doi.org/10.1007/978-3-030-58536-5_31.

[60] Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. Rdumb: A simple approach that questions our progress in continual test-time adaptation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7d640f377893fc5f22b5610e175ef7c3-Paper-Conference.pdf.

[61] Tillmann Rheude, Roland Eils, and Benjamin Wild. Cohort-based active modality acquisition. arXiv:2505.16791 [cs], December 2025. URL http://arxiv.org/abs/2505.16791.

[62] Adriel Saporta, Aahlad Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/pdf/2411.01053.

[63] Harald Semmelrock, Tony Ross-Hellauer, Simone Kopeinik, Dieter Theiler, Armin Haberl, Stefan Thalmann, and Dominik Kowald. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Mag.*, 46(2), 2025. doi: 10.1002/AAAI.70002. URL https://doi.org/10.1002/aaai.70002.

[64] Jing Shi, Yiwu Zhong, Ning Xu, Yin Li, and Chenliang Xu. A simple baseline for weakly-supervised scene graph generation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, page 16373–16382. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01608. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Shi_A_Simple_Baseline_for_Weakly-Supervised_Scene_Graph_Generation_ICCV_2021_paper.pdf.

[65] Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. arXiv:2504.07951 [cs], April 2025. URL http://arxiv.org/abs/2504.07951.

[66] 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, September 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00689-4.

[67] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard David Jean Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly Wiberg, Michael Li, Ignacio Fuentes, and Dimitris Bertsimas. Code for generating the haim multimodal dataset of mimic-iv clinical data and x-rays, 2022. URL https://physionet.org/content/haim-multimodal/1.0.1/.

[68] Andrew H. Song, Richard J. Chen, Guillaume Jaume, Anurag Jayant Vaidya, Alexander Baras, and Faisal Mahmood. Multimodal prototyping for cancer survival prediction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, page 46050–46073. PMLR, July 2024. URL https://proceedings.mlr.press/v235/song24b.html.

[69] Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NialiwI2V6.

[70] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779, March 2015. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001779.

[71] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, page 6558–6569. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1656. URL https://doi.org/10.18653/v1/p19-1656.

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, page 5998–6008, 2017.

[73] Kirill Vishniakov, Karthik Viswanathan, Aleksandr Medvedev, Praveenkumar Kanithi, Marco AF Pimentel, Ronnie Rajan, and Shadab Khan. Tokenization to transfer: Do genomic foundation models learn good representations? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4UY1NHG5Ge.

[74] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, page 15878–15887. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01524. URL https://doi.org/10.1109/CVPR52729.2023.01524.

[75] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, page 12692–12702. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01271. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_What_Makes_Training_Multi-Modal_Classification_Networks_Hard_CVPR_2020_paper.html.

[76] Yuanzhi Wang, Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Issue #2: Evaluation protocol, 2019. URL https://github.com/CheyneyComputerScience/CREMA-D/issues/2.

[77] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[78] Yuanzhi Wang, Yong Li, and Zhen Cui. Issue #5: Why before running missing cases, you should download the weights pretrained by complete multimodal data, 2024. URL https://github.com/mdswyz/IMDer/issues/5.

[79] Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning. In *2025 IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, Hawaii, October 19-23, 2025*. IEEE, 2025. URL https://arxiv.org/abs/2507.10213.

[80] Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *2025 IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, Hawaii, October 19-23, 2025*. IEEE, 2025. URL https://arxiv.org/pdf/2507.10203.

[81] Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, page 52559–52572. PMLR, July 2024. URL https://proceedings.mlr.press/v235/wei24d.html.

[82] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, page 24043–24055. PMLR, July 2022. URL https://proceedings.mlr.press/v162/wu22d.html.

[83] Qilong Wu, Yiyang Shao, Jun Wang, and Xiaobo Sun. Learning optimal multimodal information bottleneck representations. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=5TUa2UXSpp.

[84] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. arXiv:2409.07825 [cs], October 2024. URL http://arxiv.org/abs/2409.07825.

[85] Yifan Xia, Xianliang Yang, Zichuan Liu, Zhihao Liu, Lei Song, and Jiang Bian. Position: Rethinking post-hoc search-based neural approaches for solving large-scale traveling salesman problems. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=cEJ9jNJuJP.

[86] Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3dvlm: An efficient vision-language model for 3d medical image analysis. arXiv: 2503.20047, 2025. URL https://doi.org/10.48550/arXiv.2503.20047.

[87] Wanting Xu, Yang Liu, Langping He, Xucheng Huang, and Ling Jiang. Xmodel-vlm: A simple baseline for multimodal vision language model. arXiv:2405.09215 [cs], June 2024. URL http://arxiv.org/abs/2405.09215.

[88] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 3718–3727, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.343. URL https://www.aclweb.org/anthology/2020.acl-main.343.

[89] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:1606.06259 [cs], August 2016. URL http://arxiv.org/abs/1606.06259.

[90] Amir Zadeh, Paul Pu Liang, Soujanya Poria, E. Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Annual Meeting of the Association for Computational Linguistics*, 2018. URL https://aclanthology.org/P18-1208.pdf.

[91] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv:2506.05176 [cs], June 2025. URL http://arxiv.org/abs/2506.05176.

[92] Yaqian Zhang, Bernhard Pfahringer, Eibe Frank, Albert Bifet, Nick Jin Sean Lim, and Yunzhe Jia. A simple but strong baseline for online continual learning: Repeated augmented rehearsal. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/5ebbbac62b968254093023f1c95015d3-Paper-Conference.pdf.

[93] Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, page 60674–60703. PMLR, July 2024. URL https://proceedings.mlr.press/v235/zhao24b.html.

[94] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1367–1376, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Zhou_A_Simple_Baseline_for_Efficient_Hand_Mesh_Reconstruction_CVPR_2024_paper.pdf.

[95] Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. A simple baseline for spoken language to sign language translation with 3d avatars. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLIX*, volume 15107 of *Lecture Notes in Computer Science*, page 36–54. Springer, 2024. doi: 10.1007/978-3-031-72967-6_3. URL https://doi.org/10.1007/978-3-031-72967-6_3.

A Broader Impact and Ethics

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

B Method Reimplementation Details

For every method, weight initialization is applied throughout the entire architecture, including encoders and method-specific extensions (Kaiming initialization for linear and convolutional layers with biases set to zero, constant values for the weight (1.0) and bias (0.0) of LayerNorm layers, and Gaussian initialization for additional parameters). Model-related blocks*, e.g.,* for Transformers or ResNets, are not re-initialized specifically. For Transformer blocks, sinusoidal positional encodings are used. Unless a method architecturally requires full token sequences*, e.g.,* cross-attention between modality token sequences, we propagate a single modality token*, e.g.,* a `[CLS]` representation, from each encoder for efficiency and to keep bandwidth comparable across methods.

C Compute Environment

Our experiments are conducted on a High-Performance Cluster (HPC) with the following environment:

- 21 Dell PowerEdge R7525 compute nodes, each with 64 AMD Epyc cores (Rome), 512GB RAM and 1 NVIDIA A100 40G GPU
- 2 Dell PowerEdge XE8545 compute nodes, each with 128 AMD Epyc cores (Milan), 512GB RAM, 4 NVIDIA A100 40G GPUs (NVLink-connected)

All efficiency-related experiments are done on the first*, i.e.,* single-GPU compute node.

D Detailed Model Results

In addition to Figure 2, Tables 4 and 5 present the detailed model results across the datasets. For datasets without a sequence dimension, SimBaMM reduces to SimBaMM$^{\text{CLS}}$, resulting in n/a for the other SimBaMM variant.

E Detailed Efficiency Analysis

In addition to the tables representing efficiency vs. performance, Figure 6 and Table 6 present a focused view on efficiency comparisons.

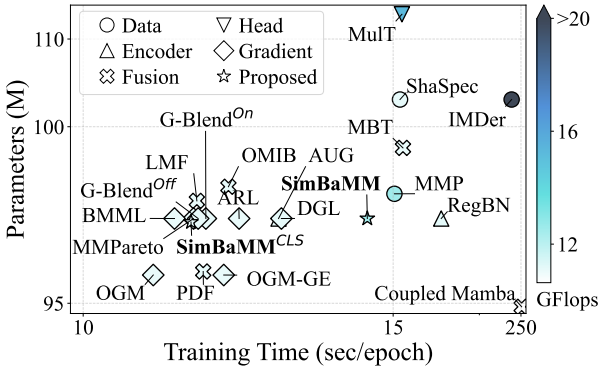


Figure 6: Visual representation of the efficiency of the methods w.r.t. the number of parameters, training time, and FLOPS.

Table 4: Performance in terms of AUROC. Details including full unimodal results in Section F. No method is highlighted because performance differences fall within the reported mean ± SDs, summarized in Table 3.

Method	HAIM ↑	Symile ↑	INSPECT ↑	UKB ↑
Best Unimodal	$0.7042 \pm .020$	$0.6453 \pm .030$	$0.6568 \pm .003$	$0.7560 \pm .010$
Missing Data-Based				
IMDer [77]	$0.6900 \pm .018$	$0.6266 \pm .016$	$0.6385 \pm .007$	$0.7969 \pm .003$
MMP [56]	$0.6611 \pm .016$	$0.5824 \pm .054$	$0.6289 \pm .018$	$0.7984 \pm .004$
ShaSpec [74]	$0.6863 \pm .019$	$0.5823 \pm .052$	$0.6348 \pm .006$	$0.7968 \pm .007$
SimBaMM^{CLS}	$0.6558 \pm .031$	$0.5925 \pm .045$	$0.6378 \pm .008$	$0.7957 \pm .008$
SimBaMM	$0.6600 \pm .024$	$0.6262 \pm .023$	n/a	n/a
Encoder-Based				
RegBN [19]	$0.6912 \pm .054$	$0.6412 \pm .018$	$0.6379 \pm .002$	$0.7919 \pm .001$
AUG [30]	$0.7103 \pm .027$	$0.6101 \pm .035$	$0.6513 \pm .016$	$0.7361 \pm .136$
Fusion-Based				
MBT [55]	$0.6804 \pm .029$	$0.6331 \pm .023$	$0.6391 \pm .003$	$0.7456 \pm .106$
LMF [46]	$0.7181 \pm .037$	$0.6146 \pm .023$	$0.6454 \pm .003$	$0.4999 \pm .000$
PDF [8]	$0.6880 \pm .025$	$0.5922 \pm .020$	$0.6419 \pm .003$	$0.7825 \pm .014$
Coupled Mamba [39]	$0.6631 \pm .040$	$0.6603 \pm .027$	$0.6580 \pm .004$	$0.7974 \pm .003$
OMIB [83]	$0.7040 \pm .019$	$0.5922 \pm .038$	$0.6272 \pm .008$	$0.6262 \pm .070$
Head-Based				
MulT [71]	$0.6872 \pm .030$	$0.6374 \pm .014$	$0.6510 \pm .004$	OOM
Gradient-Based				
OGM [57]	$0.7172 \pm .014$	$0.6138 \pm .011$	$0.6431 \pm .005$	$0.7982 \pm .013$
OGM^{GE} [57]	$0.7069 \pm .030$	$0.6152 \pm .034$	$0.6426 \pm .002$	$0.7982 \pm .008$
DGL [79]	$0.6143 \pm .066$	$0.6032 \pm .017$	$0.6276 \pm .043$	$0.5803 \pm .053$
ARL [80]	$0.7000 \pm .024$	$0.6393 \pm .032$	$0.6514 \pm .002$	$0.7986 \pm .003$
MMPareto [81]	$0.7021 \pm .029$	$0.6319 \pm .038$	$0.6559 \pm .005$	$0.8054 \pm .009$
BMML [82]	$0.7072 \pm .034$	$0.6419 \pm .020$	$0.6233 \pm .003$	$0.7940 \pm .003$
G-Blend Online [75]	$0.7123 \pm .028$	$0.6179 \pm .084$	$0.6538 \pm .005$	$0.8036 \pm .014$
G-Blend Offline [75]	$0.7148 \pm .032$	$0.6050 \pm .067$	$0.6284 \pm .018$	$0.7795 \pm .033$
SimBaMM^{CLS}	$0.6985 \pm .025$	$0.6318 \pm .015$	$0.6556 \pm .006$	$0.7957 \pm .008$
SimBaMM	$0.6871 \pm .029$	$0.6429 \pm .036$	n/a	n/a

F Dataset Details

MIMIC Symile The MIMIC Symile dataset provides a challenging clinical benchmark for multi-label diagnostic prediction. Derived from the MIMIC database, it comprises $10\,345$ samples collected from patients in intensive care units. The task is to predict the presence of ten conditions: Fracture, Enlarged Cardiomediastinum, Consolidation, Atelectasis, Edema, Cardiomegaly, Lung Lesion, Lung Opacity, Pneumonia, and Pneumothorax. Our evaluation leverages three distinct data types: laboratory test results, chest X-ray images, and electrocardiograms (ECGs). We use an MLP for laboratory tests, a ViT for vision data, and a Transformer encoder for ECGs. The ECG samples are preprocessed by removing baseline wander.

MIMIC HAIM Similarly, the MIMIC HAIM dataset is a healthcare benchmark focused on the same ten-class diagnostic prediction task as Symile. This bimodal dataset is substantially larger, containing $45\,050$ samples. For this dataset, our experiments utilize laboratory values and chest X-ray images as the two input modalities. We use an MLP for laboratory tests and a ViT for vision data.

INSPECT The INSPECT dataset [28] comprises $22\,449$ CT pulmonary angiography (CTPA) studies for pulmonary embolism (PE) diagnosis, combining 3D computer-assisted tomography (CT) images with longitudinal electronic health records (EHRs). We focus on the binary PE detection task, where approximately 20% of cases are PE-positive. We use MOTOR [69] embeddings for structured

Table 5: Performance in terms of Acc_7 (MOSI/MOSEI), Acc_5 (CH-SIMS/CH-SIMS2) and Acc_6 (Crema-D). No method is highlighted because performance differences fall within the reported mean $\pm$ SDs, summarized in Table 3.

Method	**MOSI** ↑	**MOSEI** ↑	**CH-SIMS** ↑	**CH-SIMS 2** ↑	**Crema-D** ↑
Unimodal					
Language	$0.4084 \pm .024$	$0.5144 \pm .006$	$0.5458 \pm .022$	$0.4280 \pm .014$	n/a
Vision	$0.2556 \pm .038$	$0.4285 \pm .007$	$0.3593 \pm .014$	$0.2497 \pm .010$	$0.3192 \pm .027$
Audio	$0.2401 \pm .011$	$0.4197 \pm .005$	$0.4298 \pm .021$	$0.2686 \pm .031$	$0.6081 \pm .032$
Missing Data-Based					
IMDer [77]	$0.2810 \pm .058$	$0.4273 \pm .039$	$0.4541 \pm .079$	$0.3951 \pm .022$	$0.6840 \pm .032$
MMP [56]	$0.2451 \pm .046$	$0.2760 \pm .117$	$0.4201 \pm .127$	$0.3589 \pm .024$	$0.6552 \pm .025$
ShaSpec [74]	$0.28921 \pm .031$	$0.3489 \pm .111$	$0.4806 \pm .026$	$0.4222 \pm .015$	$0.6051 \pm .034$
SimBaMM$^{\text{CLS}}$	$0.2683 \pm .057$	$0.3769 \pm .083$	$0.4716 \pm .062$	$0.4059 \pm .017$	$0.6380 \pm .039$
SimBaMM	$0.3028 \pm .060$	$0.4169 \pm .076$	$0.5122 \pm .047$	$0.4090 \pm .012$	n/a
Encoder-Based					
RegBN [19]	$0.2115 \pm .096$	$0.3894 \pm .069$	$0.4704 \pm .057$	$0.4135 \pm .034$	$0.6340 \pm .026$
AUG [30]	$0.2756 \pm .067$	$0.4395 \pm .022$	$0.4797 \pm .050$	$0.3715 \pm .025$	$0.6105 \pm .029$
Fusion-Based					
MBT [55]	$0.3124 \pm .053$	$0.4434 \pm .030$	$0.5113 \pm .058$	$0.4367 \pm .016$	$0.6525 \pm .018$
LMF [46]	$0.2842 \pm .042$	$0.3704 \pm .087$	$0.5020 \pm .051$	$0.4074 \pm .021$	$0.6363 \pm .013$
PDF [8]	$0.3315 \pm .048$	$0.4649 \pm .020$	$0.5109 \pm .042$	$0.4006 \pm .015$	$0.5976 \pm .028$
C-Mamba [39]	$0.3147 \pm .037$	$0.5188 \pm .005$	$0.5209 \pm .041$	$0.4448 \pm .010$	$0.6792 \pm .040$
OMIB [83]	$0.3211 \pm .015$	$0.4795 \pm .011$	$0.4535 \pm .075$	$0.4105 \pm .029$	$0.6690 \pm .046$
Head-Based					
MulT [71]	$0.3584 \pm .031$	$0.5147 \pm .006$	$0.4979 \pm .102$	$0.3993 \pm .100$	$0.6102 \pm .037$
Gradient-Based					
OGM [57]	$0.2883 \pm .063$	$0.3060 \pm .176$	$0.4766 \pm .052$	$0.3927 \pm .019$	$0.6185 \pm .043$
OGM$^{\text{GE}}$ [57]	$0.3147 \pm .026$	$0.4274 \pm .054$	$0.5617 \pm .031$	$0.3688 \pm .024$	$0.6281 \pm .031$
DGL [79]	$0.3533 \pm .046$	$0.3253 \pm .143$	$0.3948 \pm .053$	$0.2471 \pm .039$	$0.6717 \pm .033$
ARL [80]	$0.2574 \pm .073$	$0.2838 \pm .132$	$0.5570 \pm .013$	$0.4579 \pm .008$	$0.6318 \pm .032$
MMPareto [81]	$0.3488 \pm .067$	$0.4209 \pm .055$	$0.5373 \pm .035$	$0.4354 \pm .030$	$0.4398 \pm .046$
BMML [82]	$0.2287 \pm .056$	$0.4307 \pm .100$	$0.5514 \pm .035$	$0.4255 \pm .041$	$0.6535 \pm .067$
G-Blend On [75]	$0.2519 \pm .064$	$0.4128 \pm .066$	$0.5139 \pm .028$	$0.4503 \pm .010$	$0.5566 \pm .069$
G-Blend Off [75]	$0.2383 \pm .086$	$0.4290 \pm .066$	$0.5311 \pm .017$	$0.4101 \pm .023$	$0.6241 \pm .023$
SimBaMM$^{\text{CLS}}$	$0.3229 \pm .054$	$0.4936 \pm .015$	$0.5086 \pm .042$	$0.4351 \pm .014$	$0.6723 \pm .031$
SimBaMM	$0.3056 \pm .080$	$0.3950 \pm .101$	$0.5276 \pm .059$	$0.4326 \pm .028$	n/a

EHR data and Med3dVLM [86] embeddings (latest vision encoder stage with global average pooling on the feature maps) for the CT volumes, in combination with MLP encoders.

UKB The UKB is a large, long-term prospective biobank study with $488\,131$ samples after preprocessing and exclusions. While various targets*, e.g.,* for survival analysis are possible, we chose to predict ten-year mortality from the recruitment date for simplicity. It comprises 23 modalities*, i.e.,* metabolomics, EHR, PRS, proteomics, blood biochemistry, environment, stiffness, anthropometry, blood pressure, ECG, eye measures, bone densitometry, grip strength, spirometry, touchscreen, cognitive function, hearing tests, interviews, blood counts, urine assays, telomeres, and infectious diseases. The UKB category IDs used for the structured field extraction are: 220 (NMR metabolomics), 1838 (Proteomics), 100008 (Anthropometry), 100007 (Arterial stiffness), 100094 (Baseline characteristics), 17518 (Blood biochemistry), 100081 (Blood count), 100011 (Blood pressure), 100018 (Bone densitometry of heel), 100026 (Cognitive function), 104 (ECG at rest), 100012 (ECG during exercise), 100013 (Eye measures), 300/264/100313 (Genetics), 100019 (Hand grip strength), 100049 (Hearing test), 51428 (Infectious diseases), 113 (Local environment), 100020 (Spirometry), 265 (Telomeres), 100025 (Touchscreen), 100083 (Urine assays), and 100071 (Verbal interview). We use MLPs for all modalities and raw inputs except for the EHR modality, for which we use QWEN [91] embeddings based on previously used serializations [26]. Individual missing entries within a modality were imputed with zeros, and instances where a feature vector was entirely absent were excluded.

Table 6: Tabular representation of the efficiency of the methods w.r.t. the number of parameters, training time, and FLOPS.

Group	Method	#Parameters (M) ↓	FLOPs (G) ↓	Training Time (sec/epoch) ↓
Data	IMDer [77]	103.0	408.23	203.50
	MMP [56]	98.1	12.57	15.42
	ShaSpec [74]	103.0	11.14	17.45
Encoder	RegBN [19]	97.4	11.15	43.21
	AUG [30]	97.4	11.10	12.15
Fusion	MBT [55]	99.4	11.11	18.62
	LMF [46]	97.9	11.09	11.83
	PDF [8]	95.9	11.08	11.94
	Coupled Mamba [39]	94.9	10.63	253.80
	OMIB [83]	98.3	11.11	12.34
Head	MulT [71]	113.0	15.32	18.20
Gradient	OGM [57]	95.8	11.08	11.13
	OGM^{GE} [57]	95.8	11.08	12.27
	DGL [79]	97.4	11.14	13.20
	ARL [80]	97.4	11.10	12.52
	MMPareto [81]	97.4	11.10	11.78
	BMML [82]	97.4	11.10	11.47
	$\text{G-Blend}^{\text{Online}}$ [75]	97.4	11.10	11.98
	$\text{G-Blend}^{\text{Offline}}$ [75]	97.4	11.10	11.85
	$\text{SimBaMM}^{\text{CLS}}$	97.3	11.10	11.74
	SimBaMM	97.4	12.46	14.58

CMU-MOSI and CMU-MOSEI Shifting from the clinical domain, the CMU-MOS(E)I dataset is a benchmark for multimodal sentiment analysis and emotion recognition. It consists of 2199 (MOSI) and 22 856 (MOSEI) video clips of speakers expressing opinions, with annotations for seven emotional classes. The data is trimodal, encompassing visual, acoustic, and linguistic information. A key distinction from the other datasets is the use of the officially provided pre-computed feature embeddings, in contrast to processing raw data. We use BERT for language sequences and a Transformer encoder for the vision and audio embeddings.

CH-SIMS and CH-SIMS 2 These datasets are benchmarks for multimodal sentiment analysis and emotion recognition. They consist of 2281 (CH-SIMS) and 4403 (CH-SIMS 2) video clips of speakers expressing emotions. Similar to MOS(E)I, they are trimodal but provide raw data. We use Wav2Vec2, ViViT, and BERT for the audio, video, and language data, respectively.

Crema-D Crema-D is a dataset for multimodal sentiment analysis with six emotion classes (such as anger and fear) that can be further subdivided into a more fine-grained set of 24 classes. It consists of 7442 samples and provides audio and video data. Following related work, we use spectrograms in combination with a ResNet for the audio data and video frames in combination with another ResNet for the vision data. For the case study, we use horizontal flipping augmentations and ImageNet normalization values analogous to the AUG code repository.

G Case Study Details

Especially for the case study, alternative viewpoints on our set of implementation choices are possible. Therefore, we record key concerns, add further experiments and explain how they relate to the case study.

On the Necessity of Reusing Splits for Fair Comparison It is often argued that new methods should use legacy data splits to ensure numbers are directly comparable to previous tables. Comparability with prior results is a pragmatic concern, but we do not consider reusing legacy splits as a methodological requirement. In particular, when existing splits are underspecified or violate stated

Table 7: Unimodal performance of the modalities for the healthcare datasets. For the UKB, if unimodal data is not available for a sample, it is zero-imputed to ensure a fair comparison to the multimodal performance.

Modality	**AUROC** ↑
HAIM	
X-ray	0.7042 ± 0.020
Laboratory	0.6654 ± 0.038
Symile	
X-ray	0.6453 ± 0.030
Laboratory	0.5597 ± 0.010
ECG	0.4960 ± 0.024
INSPECT	
CT	0.5825 ± 0.007
EHR	0.6568 ± 0.003
UKB	
Metabolomics	0.7108 ± 0.003
EHR	0.7560 ± 0.010
Proteomics	0.5434 ± 0.009
Polygenic risk score (PRS)	0.6008 ± 0.008
Blood biochemistry	0.7156 ± 0.004
Baseline characteristics	0.7151 ± 0.004
Local environment	0.5049 ± 0.009
Arterial stiffness	0.5290 ± 0.016
Anthropometry	0.6086 ± 0.025
Blood pressure	0.6403 ± 0.008
ECG	0.5298 ± 0.017
Eye measures	0.5233 ± 0.014
Bone densitometry	0.5853 ± 0.003
Grip strength	0.5227 ± 0.008
Spirometry	0.6668 ± 0.021
Touchscreen	0.7334 ± 0.007
Cognitive function	0.5956 ± 0.006
Hearing	0.5355 ± 0.020
Interview	0.6803 ± 0.005
Blood count	0.6810 ± 0.005
Urine assays	0.5384 ± 0.006
Telomeres	0.5670 ± 0.010
Infectious diseases	0.5008 ± 0.001

dataset guidance*, e.g.,* subject independence, adopting a standardized, leakage-avoiding protocol is justified even if it breaks comparability with earlier reports. Rather than treating inherited splits as binding, we view them as one point in the design space: authors can report results on legacy splits for continuity while also providing results under a corrected protocol to support valid conclusions. In this sense, the case study targets evaluation practice*, i.e.,* how conclusions shift under a more defensible protocol, rather than attributing shortcomings to any single method or paper.

AUG with Linear Head We replace AUG's original linear head with a Transformer-based head to align the architectural capacity with other compared methods. To verify that this choice does not drive our conclusions, we additionally show that AUG with a linear head performs similarly (AUG-L, Figure 7). AUG shows the same overall performance trends and leads to the same conclusions about its relative behavior in the multimodal setting, indicating that the case study findings are not an artifact of the Transformer head.

Different Optimizers To assess whether our conclusions depend on the choice of optimizer, we performed an additional hyperparameter sweep using AdamW instead of the originally used SGD (AUG-AdamW, Figure 7). To circumvent learning rate schedulers, we still use the ScheduleFree variant. This change does not alter the qualitative outcome: the relative behavior of AUG and the baselines, as well as the resulting ranking trends that motivate the case study, remain consistent. This

Table 8: Unimodal performance of the modalities for the UKB without zero imputation, *i.e.,* only the available modality subset is used.

Modality	**AUROC** ↑
Metabolomics	0.7118 ± 0.004
EHR	0.7728 ± 0.005
Proteomics	0.7588 ± 0.018
PRS	0.6009 ± 0.009
Blood biochemistry	0.7183 ± 0.004
Baseline characteristics	0.7150 ± 0.006
Local environment	0.5025 ± 0.004
Arterial stiffness	0.5997 ± 0.009
Anthropometry	0.6060 ± 0.013
Blood pressure	0.6388 ± 0.009
ECG	0.6869 ± 0.005
Eye measures	0.5669 ± 0.045
Bone densitometry	0.6008 ± 0.002
Grip strength	0.5212 ± 0.008
Spirometry	0.6615 ± 0.004
Touchscreen	0.7180 ± 0.006
Cognitive function	0.5956 ± 0.005
Hearing	0.6020 ± 0.009
Interview	0.6656 ± 0.006
Blood count	0.6838 ± 0.006
Urine assays	0.5284 ± 0.009
Telomeres	0.5685 ± 0.009
Infectious diseases	0.5330 ± 0.027

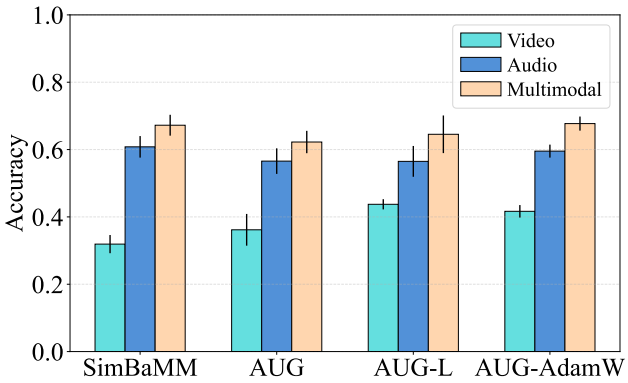


Figure 7: Alternative configurations for AUG in the case study setting with the corrected data protocol, *i.e.,* with a linear head (AUG-L) and another optimizer (AUG-AdamW).

indicates that our headline conclusions are not an artifact of selecting a particular optimizer family, but persist under a substantially different optimization setup.

Sweep on Full Parameter Space One may argue that initializing AUG's sweep from architectural choices selected for our method, *e.g.,* encoder and head dimensions, could bias the comparison, and that AUG should instead be tuned from scratch over the full design space, including architectural hyperparameters as well as optimization settings. This concern is addressed by AUG-L (Figure 7), since all model hyperparameters are sweeped in this setting, except for the encoders which are set to the originally used ResNet18s. The resulting performance and qualitative trends are unchanged: the relative conclusions of the case study and the observed method ranking remain consistent. This indicates that our findings are robust to the hyperparameter search procedure and do not arise from anchoring AUG to architectural choices optimized for a different method.

Differences to Original Results For DGL and ARL on CREMA-D, our experiments reveal the same issue identified for AUG: the reported gains rely on model selection using the test set. When applying our strictly separated validation protocol, performance decreases substantially. We reproduced this in the official code repositories as well. First, as for our case study (on AUG), the train/val/test splits contain overlapping speaker identities across training and evaluation sets. Second, checkpoints are selected using test-set performance. We replaced these with speaker-disjoint $60/20/20$ train/validation/test splits and selected checkpoints on the validation set. With these corrections, the official implementations reach decreased performance comparable to the results which we report in our paper. For DGL, this means the original result of 77.48 decreases to 67.3. For ARL, this means that the original result of 76.61 decreases to 65.9

H Statistical Comparison Across Datasets

To rigorously compare methods across datasets, we employ Bayesian hierarchical analysis following Benavoli et al. [5], Corani et al. [14]. The model uses fold-level cross-validation results, accounts for correlation between folds due to overlapping training sets, and yields direct posterior probabilities for each pairwise comparison.

H.1 Bayesian Hierarchical Pairwise Comparisons

For each pair of methods (i, j), we compare their fold-level cross-validation results across the datasets on which both methods are available. Let $x_{d,f} = s_{d,i,f} - s_{d,j,f}$ be the per-fold performance difference on dataset d. We model within-dataset differences as correlated observations due to shared training data in K-fold cross-validation:

$$\mathbf{x}_d \sim \mathcal{N}(\mu_d \mathbf{1}, \Sigma_d), \qquad \rho = 1/K, \tag{1}$$

where K is the number of folds (here $K = 5$, thus $\rho = 0.2$). Dataset-level mean differences are modeled with a robust population distribution:

$$\mu_d \sim t(\delta_0, \sigma_0, \nu), \tag{2}$$

where δ_0 is the population-level mean difference and ν represents the degrees of freedom, accounting for non-normality. We define a ROPE of $\pm\epsilon$ (here $\epsilon = 0.01$ in absolute metric points). Following Corani et al. [14], we report the posterior probabilities that each of three outcomes is the *most probable* outcome on a new dataset drawn from the population distribution $\delta_{\text{new}} \sim t(\delta_0, \sigma_0, \nu)$: *method j better* ($\delta_{\text{new}} < -\epsilon$), *practically equivalent* ($|\delta_{\text{new}}| \leq \epsilon$), and *method i better* ($\delta_{\text{new}} > \epsilon$). Concretely, for each posterior draw of $(\delta_0, \sigma_0, \nu)$ we compute the three region probabilities for δ_{new} and count which region has the largest probability; the reported values are the resulting frequencies. We complete the model with the following hyperpriors based on the robust ranges identified in Corani et al. [14]:

$$\delta_0 \sim \mathcal{U}(-1, 1), \tag{3}$$
$$\sigma_0 \sim \mathcal{U}(0, \bar{s}_0), \tag{4}$$
$$\nu \sim \text{Gamma}(\alpha, \beta) + 1, \tag{5}$$
$$\alpha \sim \mathcal{U}(1.0, 2.0), \quad \beta \sim \mathcal{U}(0.01, 0.10). \tag{6}$$

Following Corani et al. [14], we set $\bar{s}_0 = 1000\, s_{\bar{x}}$, where $s_{\bar{x}}$ is the standard deviation across datasets of the per-dataset mean differences $\bar{x}_d = \frac{1}{K} \sum_{f=1}^{K} x_{d,f}$.

Reported quantities For every comparison, we additionally report the posterior mean and 95% credible interval of the population-level mean difference δ_0 (absolute metric points; $\delta_0 > 0$ indicates method i is better on average).

Handling missingness Missing method-dataset pairs are handled via pairwise complete cases: for a comparison (i, j) we include only datasets where *both* methods have fold-level results available.

Model fitting and diagnostics We fit the model with NUTS using PyMC (nutpie backend) with four chains (2000 warmup and 4000 post-warmup draws per chain). We monitor $\hat{R}$, effective sample

Table 9: Comparison with the best unimodal baseline.

Method	P(base >other)	P(≈)	P(other >base)	$\mathbb{E}[\delta_0]$	95% CI
CoupledMamba	0.18	0.10	0.72	-0.010	[-0.038,0.024]
SimBaMM-CLS	0.41	0.06	0.53	-0.002	[-0.031,0.031]
MMPareto	0.66	0.08	0.26	0.013	[-0.026,0.066]
ARL	0.19	0.14	0.67	-0.008	[-0.034,0.024]
GBlend-On	0.61	0.10	0.29	0.011	[-0.029,0.062]
BMML	0.59	0.03	0.38	0.006	[-0.035,0.056]
OGM-GE	0.80	0.02	0.18	0.017	[-0.024,0.060]
MulT	0.29	0.67	0.04	0.006	[-0.008,0.025]
SimBaMM-Full	0.74	0.14	0.13	0.021	[-0.023,0.083]
MBT	0.91	0.03	0.06	0.021	[-0.009,0.054]
GBlend-Off	0.73	0.18	0.08	0.015	[-0.014,0.046]
LMF	0.93	0.00	0.07	0.057	[-0.020,0.142]
OMIB	0.96	0.00	0.04	0.033	[-0.007,0.075]
OGM	0.80	0.05	0.14	0.018	[-0.020,0.064]
PDF	0.95	0.03	0.03	0.023	[-0.002,0.050]
AUG	0.94	0.02	0.04	0.031	[-0.005,0.072]
RegBN	0.68	0.04	0.27	0.013	[-0.026,0.067]
DGL	0.98	0.00	0.02	0.083	[0.009,0.161]

size (ESS), divergences, and max tree depth hits for the key parameters $(\delta_0, \sigma_0, \nu)$ to ensure reliable posterior estimates. When diagnostics are poor, we automatically re-sample with a more conservative target acceptance rate and/or a larger draw budget. Table 15 summarizes these diagnostics for the baseline comparisons reported in Tables 3 and 9.

Bounded-metric sanity check Since the evaluated metrics are bounded in $[0, 1]$, the dataset-level mean difference satisfies $\delta \in [-1, 1]$. We therefore compute a simple posterior predictive check: the probability that a new dataset-level effect drawn from the fitted population distribution violates this bound, i.e., $P(|\delta_{\text{new}}| > 1)$. For the baseline comparisons in Table 15, this probability is negligible, indicating the fitted population model places essentially no mass outside the feasible range.

H.2 ROPE Sensitivity

To assess how sensitive our conclusions are to the practical significance threshold, we repeat the comparisons with $\epsilon \in \{0.005, 0.01, 0.02\}$ (i.e., ROPE $\pm$0.5%, $\pm$1%, and $\pm$2% in absolute metric points). Tables 3 and 9 correspond to $\epsilon = 0.01$; additional ROPE settings are reported below.

Key takeaway The qualitative conclusions reported in Section 4.2 are robust to reasonable ROPE choices: tightening the ROPE makes more comparisons count as "wins" by a small margin (reducing $P(\approx)$), whereas widening the ROPE shifts probability toward practical equivalence. Importantly, the posterior on δ_0 (mean and 95% credible interval) is unchanged; only the partitioning into *win / ROPE / loss* depends on ϵ.

Comparison to SimBaMM$^{\text{CLS}}$ For the strongest competitors, the interpretation changes smoothly with ϵ. For example, CoupledMamba has $P(\text{CoupledMamba} > \text{SimBaMM-CLS}) = 0.64$ under $\epsilon = 0.005$ (Table 10), but this drops to 0.30 under $\epsilon = 0.01$ (Table 3) and to 0.04 under $\epsilon = 0.02$ (Table 11), while $P(\approx)$ increases from $0.28 \rightarrow 0.67 \rightarrow 0.95$. Thus, if a $\geq 1\%$ improvement is considered practically meaningful (as in Section 4.2), no method shows strong evidence of outperforming SimBaMM-CLS; if a tighter threshold is used, CoupledMamba appears more likely to be *slightly* better, consistent with a small estimated mean effect (Table 3 reports $\mathbb{E}[\delta_0] \approx -0.006$ for SimBaMM-CLS minus CoupledMamba).

Comparison to the best unimodal baseline The sensitivity analysis suggests that improvements over the best unimodal baseline (when present) are typically modest: CoupledMamba and ARL are likely to beat the unimodal baseline for $\epsilon \in \{0.005, 0.01\}$ (e.g., $P(\text{CoupledMamba} > \text{Unimodal}) = 0.79/0.72$ and $P(\text{ARL} > \text{Unimodal}) = 0.76/0.67$ for $\epsilon = 0.005/0.01$), but these probabilities decrease to 0.37 and 0.28 at $\epsilon = 0.02$ with $P(\approx)$ becoming dominant (Table 13). This supports the

Table 10: Comparison with SimBaMM-CLS (ROPE $\pm 0.5\%$).

Method	P(base >other)	P(≈)	P(other >base)	$\mathbb{E}[\delta_0]$	95% CI
CoupledMamba	0.07	0.28	0.64	-0.006	[-0.021,0.006]
Best unimodal	0.56	0.00	0.44	0.002	[-0.031,0.031]
MMPareto	0.23	0.32	0.45	-0.002	[-0.019,0.022]
ARL	0.45	0.18	0.37	0.002	[-0.019,0.028]
GBlend-On	0.48	0.23	0.29	0.005	[-0.018,0.045]
BMML	0.79	0.00	0.21	0.013	[-0.021,0.051]
OGM-GE	0.89	0.03	0.08	0.015	[-0.009,0.042]
MulT	0.47	0.13	0.41	0.002	[-0.023,0.031]
SimBaMM-Full	0.62	0.07	0.31	0.007	[-0.038,0.056]
MBT	0.90	0.05	0.05	0.012	[-0.004,0.028]
GBlend-Off	0.95	0.01	0.05	0.021	[-0.005,0.048]
LMF	0.90	0.00	0.10	0.049	[-0.028,0.142]
OMIB	0.98	0.00	0.02	0.022	[0.003,0.043]
OGM	0.95	0.01	0.04	0.017	[-0.003,0.040]
PDF	0.99	0.00	0.01	0.020	[0.005,0.037]
AUG	0.97	0.00	0.03	0.032	[-0.000,0.066]
RegBN	0.92	0.03	0.05	0.014	[-0.004,0.039]
DGL	0.98	0.00	0.02	0.078	[0.006,0.158]

Table 11: Comparison with SimBaMM-CLS (ROPE $\pm 2\%$).

Method	P(base >other)	P(≈)	P(other >base)	$\mathbb{E}[\delta_0]$	95% CI
CoupledMamba	0.01	0.95	0.04	-0.006	[-0.021,0.006]
Best unimodal	0.25	0.53	0.22	0.002	[-0.031,0.031]
MMPareto	0.05	0.91	0.04	-0.002	[-0.019,0.022]
ARL	0.11	0.84	0.06	0.002	[-0.019,0.028]
GBlend-On	0.21	0.74	0.05	0.005	[-0.018,0.045]
BMML	0.51	0.38	0.11	0.013	[-0.021,0.051]
OGM-GE	0.38	0.59	0.03	0.015	[-0.009,0.042]
MulT	0.13	0.79	0.08	0.002	[-0.023,0.031]
SimBaMM-Full	0.30	0.57	0.14	0.007	[-0.038,0.056]
MBT	0.14	0.85	0.01	0.012	[-0.004,0.028]
GBlend-Off	0.61	0.38	0.01	0.021	[-0.005,0.048]
LMF	0.87	0.04	0.10	0.049	[-0.028,0.142]
OMIB	0.65	0.35	0.01	0.022	[0.003,0.043]
OGM	0.37	0.62	0.01	0.017	[-0.003,0.040]
PDF	0.46	0.54	0.00	0.020	[0.005,0.037]
AUG	0.88	0.11	0.02	0.032	[-0.000,0.066]
RegBN	0.29	0.71	0.01	0.014	[-0.004,0.039]
DGL	0.98	0.00	0.02	0.078	[0.006,0.158]

interpretation in Section 4.2 that any gains over unimodal baselines are not reliably large in absolute terms.

H.3 ρ and Prior-Bound Sensitivity

We follow the correlation heuristic $\rho = 1/K$ and the weakly-informative uniform scale priors recommended by Corani et al. [14]. To ensure conclusions are not driven by these modeling choices, we additionally repeat the baseline comparisons using (i) an alternative fold-correlation heuristic $\rho = 1/(K-1)$ and (ii) tighter upper bounds for the uniform scale priors. Table 14 reports the maximum absolute changes in the reported probabilities and $\mathbb{E}[\delta_0]$ under these perturbations.

Sensitivity result The sensitivity results show that conclusions are stable: across the baseline comparisons, the win/ROPE/loss probabilities change by at most 0.053 and the population-level mean effect estimates $\mathbb{E}[\delta_0]$ shift by at most 0.002 under these alternative settings (Table 14).

Table 12: Comparison with the best unimodal baseline (ROPE $\pm 0.5\%$).

Method	P(base >other)	P(≈)	P(other >base)	$\mathbb{E}[\delta_0]$	95% CI
CoupledMamba	0.20	0.02	0.79	-0.010	[-0.038,0.024]
SimBaMM-CLS	0.44	0.00	0.56	-0.002	[-0.031,0.031]
MMPareto	0.69	0.02	0.29	0.013	[-0.026,0.066]
ARL	0.23	0.02	0.76	-0.008	[-0.034,0.024]
GBlend-On	0.65	0.02	0.33	0.011	[-0.029,0.062]
BMML	0.60	0.00	0.40	0.006	[-0.035,0.056]
OGM-GE	0.81	0.00	0.19	0.017	[-0.024,0.060]
MulT	0.63	0.28	0.09	0.006	[-0.008,0.025]
SimBaMM-Full	0.80	0.04	0.16	0.021	[-0.023,0.083]
MBT	0.94	0.00	0.06	0.021	[-0.009,0.054]
GBlend-Off	0.84	0.04	0.12	0.015	[-0.014,0.046]
LMF	0.93	0.00	0.07	0.057	[-0.020,0.142]
OMIB	0.96	0.00	0.04	0.033	[-0.007,0.075]
OGM	0.84	0.01	0.16	0.018	[-0.020,0.064]
PDF	0.97	0.00	0.03	0.023	[-0.002,0.050]
AUG	0.95	0.01	0.04	0.031	[-0.005,0.072]
RegBN	0.71	0.00	0.29	0.013	[-0.026,0.067]
DGL	0.98	0.00	0.02	0.083	[0.009,0.161]

Table 13: Comparison with the best unimodal baseline (ROPE $\pm 2\%$).

Method	P(base >other)	P(≈)	P(other >base)	$\mathbb{E}[\delta_0]$	95% CI
CoupledMamba	0.10	0.52	0.37	-0.010	[-0.038,0.024]
SimBaMM-CLS	0.22	0.53	0.25	-0.002	[-0.031,0.031]
MMPareto	0.52	0.34	0.14	0.013	[-0.026,0.066]
ARL	0.09	0.63	0.28	-0.008	[-0.034,0.024]
GBlend-On	0.46	0.38	0.16	0.011	[-0.029,0.062]
BMML	0.43	0.34	0.23	0.006	[-0.035,0.056]
OGM-GE	0.66	0.22	0.12	0.017	[-0.024,0.060]
MulT	0.07	0.92	0.01	0.006	[-0.008,0.025]
SimBaMM-Full	0.54	0.39	0.07	0.021	[-0.023,0.083]
MBT	0.61	0.36	0.03	0.021	[-0.009,0.054]
GBlend-Off	0.40	0.56	0.03	0.015	[-0.014,0.046]
LMF	0.93	0.01	0.07	0.057	[-0.020,0.142]
OMIB	0.90	0.06	0.04	0.033	[-0.007,0.075]
OGM	0.58	0.33	0.08	0.018	[-0.020,0.064]
PDF	0.68	0.31	0.01	0.023	[-0.002,0.050]
AUG	0.83	0.15	0.02	0.031	[-0.005,0.072]
RegBN	0.51	0.35	0.15	0.013	[-0.026,0.067]
DGL	0.98	0.00	0.02	0.083	[0.009,0.161]

Sampler diagnostics We summarize standard MCMC diagnostics for the baseline comparisons (Table 15): the maximum split-$\hat{R}$ across the key parameters $(\delta_0, \sigma_0, \nu)$, the minimum bulk and tail effective sample sizes (ESS), and whether any fits exhibit NUTS divergences or max-treedepth hits. We also report the maximum posterior predictive probability $P(|\delta_{\text{new}}| > 1)$ as a bounded-metric sanity check.

Diagnostics result The sampler diagnostics indicate adequate convergence and effective sampling for the baseline comparisons: split-$\hat{R}$ is close to 1 (max 1.01), ESS values are above common thresholds (min bulk/tail ESS 439/108), and we observe no divergences or max-treedepth issues (Table 15).

I Additional UKB Experiments

We include results on the UKB with intermediate modality counts in Table 16. Our results suggest that data scale may play an important role in realizing the benefits of multimodal learning. Multimodal

Table 14: Sensitivity of baseline comparisons to ρ and prior bounds.

Baseline	Metric	$\rho = 1/(K-1)$	$\bar{s} \times 100$
SimBaMM-CLS	max $\lvert\Delta P\rvert$ (win)	0.034	0.015
	max $\lvert\Delta P\rvert$ ($\approx$)	0.053	0.024
	max $\lvert\Delta P\rvert$ (loss)	0.024	0.024
	max $\lvert\Delta\mathbb{E}[\delta_0]\rvert$	0.001	0.001
Best unimodal	max $\lvert\Delta P\rvert$ (win)	0.036	0.012
	max $\lvert\Delta P\rvert$ ($\approx$)	0.035	0.008
	max $\lvert\Delta P\rvert$ (loss)	0.023	0.010
	max $\lvert\Delta\mathbb{E}[\delta_0]\rvert$	0.002	0.001

Table 15: Sampler diagnostics for baseline comparisons.

Baseline	max $\hat{R}$	min ESS (bulk/tail)	div /td	max $P(\lvert\delta_{\text{new}}\rvert > 1)$
SimBaMM-CLS	1.010	439/108	0/0	7.07e-04
Best unimodal	1.010	501/320	0/0	3.86e-04

models on the UKB show gains over unimodal baselines, and additional experiments with eight (PRS, Local environment, ECG, Spirometry, Bone densitometry, Cognitive function, Eye measures, Urine assays) and 16 (M=8 and Metabolomics, Blood pressure, Anthropometry, Hearing, Interview, Blood count, Telomeres, Infectious diseases) modalities further illustrate how performance trends evolve as the number of modalities increases. For MulT, computational cost scales quadratically with the number of modalities due to pairwise cross-modal attention. Hyperparameter tuning was therefore computationally infeasible for the 8- and 16-modality settings, for which we report n/a.

J Hyperparameter Search

We use Bayesian optimization minimizing the validation loss without sweeping the batch size [20]. In the following, we list the hyperparameter search space and configurations structured by the dataset and the evaluated methods. The former refers to the encoders used, while the latter refers primarily to the remaining groups. The initial SimBaMM is swept with 500 runs while the other methods are swept with an additional 100 runs.

J.1 Encoders Parameter Space

The parameter spaces for the individual datasets are shown in Listings 1 to 9.

```
# Encoders
vision.vit.vit:
    values: ["vit_b_16", "vit_b_32"]
vision.vit.dropout:
    values: [0, 0.1, 0.2]
sequential.transformer.num_hidden_layers:
    values: [2, 4, 6, 8]
sequential.transformer.num_attention_heads:
    values: [2, 4, 8, 16]
sequential.transformer.intermediate_size:
    values: [256, 512, 1024, 2048]
sequential.transformer.dropout:
    values: [0, 0.1, 0.2]
```

Listing 1: Hyperparameter search space configuration for MIMIC HAIM.

Table 16: Performance for eight and 16 modalities on the UKB.

Method	M8 **AUROC** ↑	M16 **AUROC** ↑
Best Unimodal	0.6869 ± 0.005 (ECG)	0.7118 ± 0.004 (Metabolomics)
Missing Data-Based		
IMDer [77]	0.7187 ± 0.003	0.7689 ± 0.004
MMP [56]	0.7203 ± 0.001	0.7716 ± 0.002
ShaSpec [74]	0.6831 ± 0.015	0.7640 ± 0.013
SimBaMM$^{\text{CLS}}$	0.7211 ± 0.003	0.7660 ± 0.003
SimBaMM	n/a	n/a
Encoder-Based		
RegBN [19]	0.7180 ± 0.008	0.7661 ± 0.002
AUG [30]	0.6572 ± 0.010	0.7582 ± 0.007
Fusion-Based		
MBT [55]	0.7180 ± 0.008	0.7511 ± 0.005
LMF [46]	0.5069 ± 0.004	0.4997 ± 0.001
PDF [8]	0.7033 ± 0.005	0.7646 ± 0.004
Coupled Mamba [39]	0.7182 ± 0.003	0.7272 ± 0.003
OMIB [83]	0.6752 ± 0.008	0.7272 ± 0.003
Head-Based		
MulT [71]	n/a	n/a
Gradient-Based		
OGM [57]	0.7246 ± 0.009	0.7686 ± 0.003
OGM$^{\text{GE}}$ [57]	0.6919 ± 0.009	0.7576 ± 0.002
DGL [79]	0.6364 ± 0.006	0.7165 ± 0.017
ARL [80]	0.7152 ± 0.004	0.7703 ± 0.004
MMPareto [81]	0.7195 ± 0.002	0.7678 ± 0.003
BMML [82]	0.7204 ± 0.001	0.7762 ± 0.003
G-Blend Online [75]	0.7212 ± 0.009	0.7740 ± 0.010
G-Blend Offline [75]	0.7083 ± 0.021	0.7625 ± 0.010
SimBaMM$^{\text{CLS}}$	0.7211 ± 0.003	0.7660 ± 0.003
SimBaMM	n/a	n/a

J.2 Methods Parameter Space

The parameter spaces for the individual methods are shown in Listings 10 to 27.

```
# Encoders
vision.vit.vit:
    values: ["vit_b_16", "vit_b_32"]
vision.vit.dropout:
    values: [0, 0.1, 0.2]
sequential1.transformer.num_hidden_layers:
    values: [2, 4, 6, 8]
sequential1.transformer.num_attention_heads:
    values: [2, 4, 8, 16]
sequential1.transformer.intermediate_size:
    values: [256, 512, 1024, 2048]
sequential1.transformer.dropout:
    values: [0, 0.1, 0.2]
sequential2.mlp.hidden_dims:
    values: [[256, 512, 256], [512, 1024, 512],
             [1024, 2048, 1024]]
sequential2.mlp.hidden_dropouts:
    values: [[0, 0, 0], [0.1, 0.1, 0.1],
             [0.2, 0.2, 0.2]]
```

Listing 2: Hyperparameter search space configuration for MIMIC Symile.

```
# Encoders
vision.mlp.hidden_dims:
    values: [[1024, 512, 256], [512, 256, 128],
             [256, 128, 64], [128, 64, 32]]
vision.mlp.hidden_dropouts:
    values: [[0.0, 0.0, 0.0], [0.1, 0.1, 0.1],
             [0.2, 0.2, 0.2]]
ehr.mlp.hidden_dims:
    values: [[1024, 512, 256], [512, 256, 128],
             [256, 128, 64], [128, 64, 32]]
ehr.mlp.hidden_dropouts:
    values: [[0.0, 0.0, 0.0], [0.1, 0.1, 0.1],
             [0.2, 0.2, 0.2]]
```

Listing 3: Hyperparameter search space configuration for INSPECT.

```
# Encoders, for all 23 modalities, exemplary:
nmr.mlp.hidden_dims:
    values: [[1024, 512, 256], [512, 256, 128],
             [256, 128, 64], [128, 64, 32]]
nmr.mlp.hidden_dropouts:
    values: [[0.0, 0.0, 0.0], [0.2, 0.2, 0.2],
             [0.3, 0.3, 0.3], [0.4, 0.4, 0.4],
             [0.5, 0.5, 0.5], [0.6, 0.6, 0.6],
             [0.7, 0.7, 0.7]]
```

Listing 4: Hyperparameter search space configuration for the UKB.

```
# Encoders
audio.transformer.num_hidden_layers:
    values: [1, 4, 8, 16]
audio.transformer.num_attention_heads:
    values: [1, 2, 4, 8]
audio.transformer.intermediate_size:
    values: [256, 512, 1024]
audio.transformer.dropout:
    values: [0.0, 0.1]
vision.transformer.num_hidden_layers:
    values: [1, 4, 8, 16]
vision.transformer.num_attention_heads:
    values: [1, 2, 4, 8]
vision.transformer.intermediate_size:
    values: [256, 512, 1024]
vision.transformer.dropout:
    values: [0.0, 0.1]
```

Listing 5: Hyperparameter search space configuration for MOSI.

```
# Encoders
audio.transformer.num_hidden_layers:
    values: [1, 4, 8, 16]
audio.transformer.num_attention_heads:
    values: [1, 2, 4, 8]
audio.transformer.intermediate_size:
    values: [256, 512, 1024]
audio.transformer.dropout:
    values: [0.0, 0.1]
vision.transformer.num_hidden_layers:
    values: [1, 4, 8, 16]
vision.transformer.num_attention_heads:
    values: [1, 2, 4, 8]
vision.transformer.intermediate_size:
    values: [256, 512, 1024]
vision.transformer.dropout:
    values: [0.0, 0.1]
```

Listing 6: Hyperparameter search space configuration for MOSEI.

```
# Encoders
audio.transformer.num_hidden_layers:
    values: [1, 4, 8, 16]
audio.transformer.num_attention_heads:
    values: [1, 2, 4, 8]
audio.transformer.intermediate_size:
    values: [256, 512, 1024]
vision.transformer.num_hidden_layers:
    values: [1, 4, 8, 16]
vision.transformer.num_attention_heads:
    values: [1, 2, 4, 8]
vision.transformer.intermediate_size:
    values: [256, 512, 1024]
```

Listing 7: Hyperparameter search space configuration for CHSIMS.

```
# Encoders
audio.transformer.num_hidden_layers:
    values: [1, 4, 8, 16]
audio.transformer.num_attention_heads:
    values: [1, 2, 4, 8]
audio.transformer.intermediate_size:
    values: [256, 512, 1024]
vision.transformer.num_hidden_layers:
    values: [1, 4, 8, 16]
vision.transformer.num_attention_heads:
    values: [1, 2, 4, 8]
vision.transformer.intermediate_size:
    values: [256, 512, 1024]
```

Listing 8: Hyperparameter search space configuration for CHSIMS2.

```
# Encoders
# Standard PyTorch ResNet18s
```

Listing 9: Hyperparameter search space configuration for CREMA-D.

```
# Head
head_transformer.d_model:
    values: [32, 64, 128, 256, 512]
head_transformer.dim_feedforward:
    values: [256, 512, 1024, 2048]
head_transformer.dropout:
    values: [0, 0.1, 0.2]
head_transformer.nhead:
    values: [4, 8, 16]
head_transformer.num_layers:
    values: [2, 4, 6, 8]
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
```

Listing 10: Hyperparameter search space configuration for SimBaMM.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# IMDer
imder.beta:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
ddpms.n_steps:
    values: [10, 20, 30, 40, 50, 100]
ddpms.d_model: # fixed w.r.t. Transformer d_model
    value: 256
ddpms.hidden_dim:
    values: [32, 64, 128, 256, 512, 1024]
ddpms.num_layers:
    values: [1, 2, 3, 4, 5, 6, 7, 8]
ddpms.dropout:
    values: [0.0, 0.1, 0.2]
ddpms.nhead:
    values: [1, 2, 4, 8, 16, 32]
```

Listing 11: Hyperparameter search space configuration for IMDer.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# MMP
mmp.proj_mlp.hidden_dim:
    values: [128, 256, 512, 1024]
mmp.proj_mlp.dropout:
    values: [0.0, 0.1, 0.2]
mmp.attn_steps.dropout:
    values: [0.0, 0.1, 0.2]
mmp.attn_steps.nhead:
    values: [1, 2, 4, 8, 16]
mmp.num_aggregated_tokens:
    values: [2, 4, 8, 16, 32]
mmp.loss_alignment_alpha:
    min: 0.001
    max: 1.0
    distribution: "log_uniform_values"
```

Listing 12: Hyperparameter search space configuration for MMP.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# ShaSpec
shaspec.loss_alpha:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
shaspec.loss_beta:
    min: 0.001
    max: 0.1
    distribution: "log_uniform_values"
```

Listing 13: Hyperparameter search space configuration for ShaSpec.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# RegBN
rbn.reference_modality_idx:  # 0-(M-1)
    values: [0, 1, 2]
rbn.momentum:
    values: [0.1, 0.01, 0.05]
rbn.sigma_THR:
    values: [0.1, 0.3, 0.5]
rbn.sigma_MIN:
    values: [0.01, 0.1, 0.2]
rbn.affine:
    values: [True, False]
```

Listing 14: Hyperparameter search space configuration for RegBN.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# AUG
aug.merge_alphas:
    values: [[0.33,0.33,0.33]]  # 1/M
aug.check_interval:
    values: [1, 5, 10, 20]
aug.threshold:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
aug.confidence_coeff:
    min: 0.01
    max: 10.0
    distribution: "log_uniform_values"
aug.lambda_smooth:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
```

Listing 15: Hyperparameter search space configuration for AUG.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# MBT
bottleneck.num_bottlenecks:
    values: [1, 2, 4, 8, 16, 32]
bottleneck.layers:
    values: [1, 2, 4, 8, 16, 32]
bottleneck.dim_feedforward:
    values: [128, 256, 512, 1024, 2048]
bottleneck.nhead:
    values: [1, 2, 4, 8, 16, 32]
bottleneck.dropout:
    values: [0.0, 0.1, 0.2, 0.3, 0.4]
```

Listing 16: Hyperparameter search space configuration for MBT.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# LMF
low_rank_matrix_fusion.rank:
    values: [2, 10, 20, 50, 100, 200, 250]
```

Listing 17: Hyperparameter search space configuration for LMF.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# PDF
pdf.loss_weight:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
pdf.unimodal_loss_weight:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
pdf.p_head.hidden_dims:
    values: [[128], [256], [512], [1024],
             [128, 256], [128, 256, 512],
             [128, 256, 512, 1024]]
pdf.p_head.dropout:
    values: [0.0, 0.1, 0.2]
```

Listing 18: Hyperparameter search space configuration for PDF.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# Coupled Mamba
coupled_ssm.d_model:
    values: [32, 64, 128, 256]
coupled_ssm.num_layers:
    values: [2, 3, 4, 5, 6, 7, 8]
coupled_ssm.d_state:
    values: [8, 16, 32, 64, 128]
coupled_ssm.d_conv:
    values: [2, 4, 8, 16]
coupled_ssm.expand:
    values: [1, 2, 4, 8]
```

Listing 19: Hyperparameter search space configuration for Coupled Mamba.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# OMIB
omib.beta:
    min: 0.00001
    max: 0.1
    distribution: "log_uniform_values"
omib.cross_attn_network.num_layers:
    values: [1, 2, 4, 8, 16]
omib.cross_attn_network.dropout:
    values: [0.0, 0.1, 0.2]
omib.cross_attn_network.num_heads:
    values: [1, 2, 4, 8, 16]
omib.cross_attn_network.dim_feedforward:
    values: [128, 256, 512, 1024]
omib.warmup_epochs:
    values: [0, 1, 2, 4, 8]
```

Listing 20: Hyperparameter search space configuration for OMIB.

```yaml
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# MulT
crossmodal_transformer.d_model:
    values: [256]  # = head_transformer.d_model
crossmodal_transformer.num_layers:
    values: [1, 2, 4, 8, 16]
crossmodal_transformer.dim_feedforward:
    values: [128, 256, 512, 1024]
crossmodal_transformer.nhead:
    values: [1, 2, 4, 8, 16]
crossmodal_transformer.dropout:
    values: [0.0, 0.1, 0.2, 0.5]
```

Listing 21: Hyperparameter search space configuration for MulT.

```yaml
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# OGM
ogm.alpha:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
ogm.use_ge:
    values: [False]  # True for OGM-GE
```

Listing 22: Hyperparameter search space configuration for OGM and OGM-GE.

```yaml
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# DGL
dgl.unimodal_loss_weight:
    min: 0.001
    max: 1.0
    distribution: "log_uniform_values"
```

Listing 23: Hyperparameter search space configuration for DGL.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# ARL
arl.unimodal_loss_weight:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
arl.temperature:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
```

Listing 24: Hyperparameter search space configuration for ARL.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# MMPareto
mmpareto.unimodal_loss_weight:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
mmpareto.gamma:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
```

Listing 25: Hyperparameter search space configuration for MMPareto.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# BMML
bmml.unimodal_loss_weight:
    min: 0.01
    max: 1.0
    distribution: "log_uniform_values"
bmml.alpha:
    min: 0.001
    max: 0.1
    distribution: "log_uniform_values"
bmml.q:
    values: [3, 5, 10]
bmml.bmml_momentum:
    values: [0.9, 0.99, 0.999]
bmml.warmup_epochs:
    values: [0, 1, 2, 5, 10]
```

Listing 26: Hyperparameter search space configuration for BMML.

```
# Optimizer
optimizer.lr:
    min: 0.000001
    max: 0.1
    distribution: "log_uniform_values"
optimizer.warmup_steps:
    values: [0, 100, 200, 500, 1000]
optimizer.weight_decay:
    values: [0, 0.1, 0.01, 0.001]
# GBlend
gblend.mode:
    values: ['offline']  # Online
gblend.update_freq:
    values: [1, 5, 10]
gblend.lookahead_epochs:
    values: [1, 3, 5]
```

Listing 27: Hyperparameter search space configuration for G-Blend Online/Offline.

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 main claims are summarized in Section 1. They are processed in the respective later sections, *e.g.,* Section 3 and Section 4. If not in the main paper, evidence is moved to the Appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are explicitly discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: For the theoretical result, the full set of assumptions and related proofs or derivations are provided.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

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: All the information needed to reproduce the experimental results are explained in the main paper (Section 4) and detailed in the Appendix*, e.g.,* Section J.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: We provide the code in our GitHub repository.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the training and test details in the main paper and the Appendix, *e.g.,* in Sections 4 and J.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, we provide mean results in combination with error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **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 information about the computer resources in Section C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **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: Yes, we follow the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss possible societal impacts explicitly (after Acknowledgment, before References).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not see any high risk for misuse of our work in case of these examples.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **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: [Yes]

    Justification: Yes, we credit the creators accordingly.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release our code repository with documentation and quick starts.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **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: [N/A]

    Justification: We do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **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: [N/A]

    Justification: We do not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not involved as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.