One-Versus-Others Attention: Scalable Multimodal Integration for Biomedical Data

Michal Golovanevsky [1] **Eva Schiller** [1] **Akira Nair** [1] **Ritambhara Singh** [1 2] **Carsten Eickhoff** [3]

Abstract

Multimodal models have become increasingly important as they surpass single-modality approaches on diverse tasks ranging from question-answering to autonomous driving. Despite the importance of multimodal learning, existing efforts focus on vision-language applications, where the number of modalities rarely exceeds four (images, text, audio, video). However, data in other domains, such as healthcare, may include many more modalities like X-rays, PET scans, MRIs, genetic screening, genomic data, and clinical notes, creating a need for both efficient and accurate data integration. Many multimodal foundation models rely on cross-attention or self-attention for effective data integration, which do not scale well for applications with more than two modalities. The complexity per layer of computing attention in either paradigm is, at best, quadratic with respect to the number of modalities, posing a computational bottleneck that impedes broad adoption. To address this, we propose a new attention mechanism, One-Versus-Others (OvO) attention, that scales *linearly* with the number of modalities, thus offering a significant reduction in computational complexity compared to existing multimodal attention methods. Using three biomedical datasets with diverse modalities, we show that our method decreases computation costs while increasing performance compared to popular integration techniques. Across all datasets, OvO reduced the number of required floating point operations (FLOPs) by at least 91.98%, demonstrating its significant impact on efficiency and enabling wider adaptation. [1].

1. Introduction

Multimodal learning has emerged as a promising approach, enabling joint learning from multiple data modalities (e.g., text and images). This allows for a more comprehensive and accurate understanding of tasks such as clinical decision support (Ming et al., 2022; Golovanevsky et al., 2022; Hayat et al., 2022), image and video captioning (Yu et al., 2019; Seo et al., 2022), and sentiment analysis (Poria et al., 2018). Multimodal learning has been explored through various methods in machine learning and deep learning. In deep learning, Neural Networks facilitate both intermediate fusion at any layer and late fusion at the decision-making stage. However, these fusion paradigms often lack explicit interaction between modalities. For instance, in detecting Alzheimer's Disease, genetic features reinforce clinical information, leading to more robust decision-making (Golovanevsky et al., 2022). Such interactions can be captured through the attention mechanism. Multimodal Foundation Models, such as LXMERT (Tan & Bansal, 2019) and ViLBERT (Lu et al., 2019), use cross-attention, while models like VisualBERT (Li et al., 2019) and VL-BERT (Su et al., 2019) use self-attention, as a vehicle to capture interactions between modalities.

However, both self-attention and cross-attention grow quadratically in computational burden with the number of modalities, posing a scalability challenge. In vision-language tasks, the number of modalities rarely exceeds four (e.g., images, text, video, audio), but significant bottlenecks can arise in domains like healthcare, where tasks often involve integrating data from multiple sources such as radiology, pathology, genomics, genetics, and clinical data. Thus, using cross-attention or self-attention becomes computationally impractical with many modalities.

To address this, we propose One-Versus-Others (OvO) attention, which compares one modality against a combined representation of all others through attention, reducing computational complexity to linear growth with the number of modalities (see Section 3.2). Figure 1 illustrates the difference between our approach (scaling linearly) and self-attention/cross-attention (scaling quadratically). OvO can be seamlessly integrated into existing multimodal foundation models, replacing traditional cross-attention or self-attention mechanisms.

We first present a complexity analysis and demonstrate its validity through a simulated dataset, showing scalability

[1] Code is available at `https://github.com/rsinghlab/OvO`

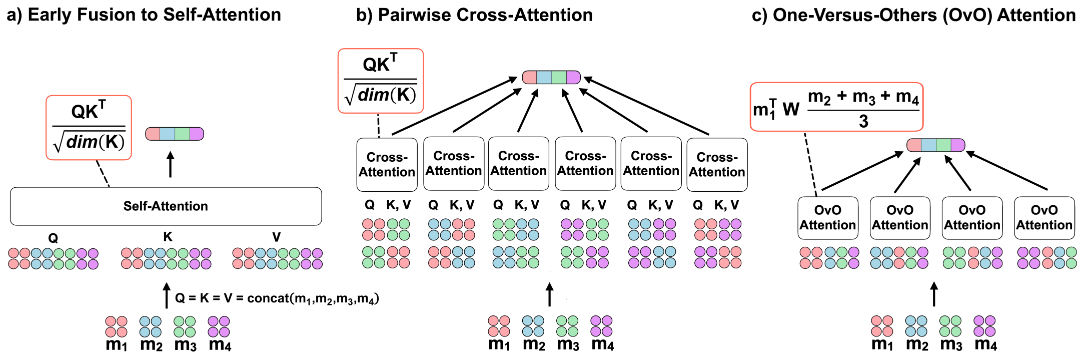
a) Early Fusion to Self-Attention
$\frac{QK^T}{\sqrt{dim(K)}}$
Self-Attention
Q
K
V
Q = K = V = concat(m1,m2,m3,m4)
m1 m2 m3 m4
b) Pairwise Cross-Attention
$\frac{QK^T}{\sqrt{dim(K)}}$
Cross-Attention
Q K, V
m1 m2 m3 m4
c) One-Versus-Others (OvO) Attention
$m_1^T W \frac{m_2 + m_3 + m_4}{3}$
OvO Attention
m1 m2 m3 m4

Figure 1. **Integration scheme comparison**. (a) Early fusion to self-attention with scaled dot product attention (Vaswani et al., 2017), and (b) Pairwise cross-attention integration with scaled dot product attention (Vaswani et al., 2017). (c) Our proposed method, One-Versus-Others (OvO), does not rely on pairwise interactions or long concatenated sequences but rather captures all modalities in a single attention score. A modality embedding is represented by m_i and W is a learnable parameter (see Section 3.1).

gains in an extreme multimodal setting (n=20). We then use 3 diverse biomedical datasets with varying modalities, encoder types, number of samples, and tasks to show improved scalability in different clinical settings. Our method reduces computation costs by at least 91.98% compared to self-attention and cross-attention, while exceeding performance.

Overall, OvO is a novel, domain-agnostic attention scheme for multimodal integration, scaling linearly with the number of modalities. It enables the practical application of deep learning models in healthcare, where computational efficiency and accuracy are crucial.

2. Related work

Multimodal attention-based models are pivotal in clinical decision support systems and vision-language applications. In the medical domain, these models are used for tasks like cancer classification (LI et al., 2021), biomarker discovery (Braman et al., 2021; Ilyin et al., 2004), and prognosis prediction (Schulz et al., 2021; Silva & Rohr, 2020), demonstrating their utility with complex medical data. Attention mechanisms serve as core components, measuring similarity among individual representations, such as word or modality-specific embeddings. Each input embedding can serve as a Query (Q), Key (K), or Value (V).

Multimodal models typically use early fusion with self-attention or cross-attention. Multimodal foundation models that use early fusion (e.g., Uniter (Chen et al., 2020), VisualBERT (Li et al., 2019), Vl-BERT (Su et al., 2019)) concatenate visual and textual embeddings before passing them through attention (see Figure 1 (a)). Given modalities m_1 and m_2, queries (Q), keys (K), and values (V) are computed from their concatenated sequence (e.g., $Q_{1,2} = concat(m_1, m_2)$). The final output from a standard Transformer block is denoted by Z, as shown in Equation 1.

$$\begin{cases} \mathrm{Z}_{1,2} = \mathit{Multiheaded\ Attention}\left(\mathrm{Q}_{1,2}, \mathrm{K}_{1,2}, \mathrm{V}_{1,2}\right) \\ \mathrm{Z} = \mathit{Transformer}(\mathrm{Z}_{1,2}) \end{cases} \tag{1}$$

Multimodal foundation models that use cross-attention (e.g., ViLBERT (Lu et al., 2019), LXMERT (Tan & Bansal, 2019), ActBERT (Zhu & Yang, 2020), MulT (Tsai et al., 2019)) input each modality into its own Transformer and then feed the outputs to a cross-modal Transformer (see Figure 1 (b)). Cross-attention captures interactions pairwise, with queries (Q), keys (K), and values (V) computed from modality inputs (m_1 and m_2). The output, Z, is shown in Equation 2.

$$\begin{cases} \mathrm{Z}_1 = \mathit{Multiheaded\ Attention}\left(\mathrm{Q}_2, \mathrm{K}_1, \mathrm{V}_1\right) \\ \mathrm{Z}_2 = \mathit{Multiheaded\ Attention}\left(\mathrm{Q}_1, \mathrm{K}_2, \mathrm{V}_2\right) \\ \mathrm{Z} = \mathit{Transformer}\left(\mathit{concat}\left(\mathrm{Z}_1, \mathrm{Z}_2\right)\right) \end{cases} \tag{2}$$

While early fusion and cross-attention can extend to three modalities (e.g., TriBERT (Rahman et al., 2021) and VATT (Akbari et al., 2021)), they face scalability challenges beyond this. Cross-attention's pairwise computations and early fusion's concatenation before attention both scale poorly with the number of modalities, increasing computational complexity quadratically (see Section 3.2). Our integration method, OvO, addresses these limitations in a scalable and domain-agnostic manner.

3. Methods

3.1. One-Versus-Others (OvO) attention

We propose a new attention mechanism, One-Versus-Others (OvO) Attention, which grows linearly with the number of modalities rather than quadratically, as is the case for cross-attention or self-attention (see Section 3.2). OvO computes attention between one modality and all other modalities. Given modality m_i from a dedicated encoder, where k is the number of modalities and $i \in 1, 2, \dots, k$, OvO takes in one modality and computes the dot product against all the other modalities with a weight matrix W. W is a learnable parameter shared across all modalities (see Figure 1 (c)) and learns interactions throughout training. The similarity score function, representing the alignment between the chosen modality and others, for modality m_i with respect to a set of other modalities $(m_j : j \neq i)$ is shown in Equation 3. The context vector in OvO for modality m_i, combining information from the other modalities, is shown in Equation 4:

$$score\,(m_i, \{m_j : \; j \; \neq \; i\}) = m_i^T \; W \; \frac{\sum_{j \, \neq \, i}^{k} \; m_j}{k-1} \qquad (3)$$

$$OvO\,(m_i, \{m_j : \; j \; \neq \; i\}) = \\ softmax(score\,(m_i, \{m_j : \; j \; \neq \; i\})) \cdot m_i \qquad (4)$$

We sum over the "other" modalities instead of concatenation for two reasons: (1) concatenation increases vector length with the number of modalities, resulting in a less scalable framework; (2) concatenation is not invariant to the order of modalities, which could affect model prediction, whereas a sum provides position invariance. Furthermore, we extend OvO attention to the multi-headed attention framework to directly compare with self-attention and cross-attention as done in (Vaswani et al., 2017), see Appendix A.4.

3.2. Model Complexity

This section highlights the complexities of early fusion followed by self-attention, pairwise cross-attention, and our One-Versus-Others (OvO) attention. Table 1 summarizes the complexity per layer. Let k represent the number of modalities, n the feature length of each modality (assuming equal), and d the representation dimension of the weight matrices. As established in (Vaswani et al., 2017), self-attention has a complexity of $\mathcal{O}(n^2 \cdot d)$. In multimodal cases, self-attention concatenates modalities before attention, leading to a sequence length of $k \cdot n$, resulting in a complexity of $\mathcal{O}(k^2 \cdot n^2 \cdot d)$. Cross-attention computes attention over all pairwise permutations of modalities: ${}_kP_2 = k(k-1)$. Thus, its complexity is $\mathcal{O}(k^2 \cdot n^2 \cdot d)$. OvO Attention requires one attention calculation per modality, making it linear with respect to k. Thus, the complexity per layer for OvO is $\mathcal{O}(k \cdot n^2 \cdot d)$. Appendix Section A provides step-by-step details for the complexity calculations.

Table 1. **Per-layer complexities of model paradigms.**

Model	Complexity Per Layer
Self-Attention	$\mathcal{O}(k^2 \cdot n^2 \cdot d)$
Cross-Attention	$\mathcal{O}(k^2 \cdot n^2 \cdot d)$
One-Versus-Others (OvO) Attention	$\mathcal{O}(k \cdot n^2 \cdot d)$

3.3. Illustration through Simulation

To illustrate the linearity of OvO compared to other paradigms, we simulated 20 artificial modalities. We created two classes: (1) 20 random feature values summing to 1.0, and (2) 20 random feature values each less than 0.15. For more details on how the threshold was chosen, see Appendix B. Using 2, 5, 10, 15, and 20 simulated modalities, we examine computation costs across the three integration methods. Notably, while self-attention and cross-attention grow quadratically with respect to the number of modalities k ($\mathcal{O}(k^2 \cdot n^2 \cdot d)$), our method scales linearly ($\mathcal{O}(k \cdot n^2 \cdot d)$), as shown in Figure 2.

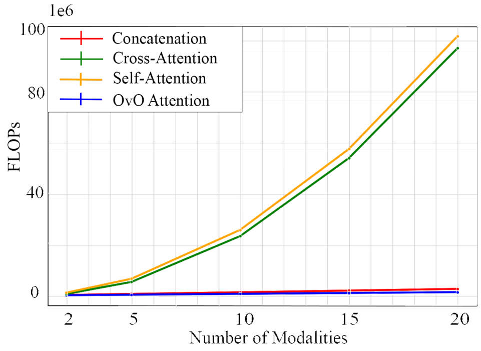

Figure 2. **Impact of using OvO attention to fuse simulated data**.

4. Experiments

We used three diverse biomedical datasets to examine our method against three standard integration techniques: concatenation with no attention (baseline), early fusion with self-attention, and pairwise cross-attention. These tasks feature a range of rich modalities that, despite their high integration costs, remain essential to solve. For implementation details and hyperparameters, see Appendix D.

4.1. Dataset descriptions

The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge (Marinescu et al., 2019) provides six modalities: cognitive tests, MRI ROIs, FDG PET ROI averages, AV45 PET ROI averages, demographic information, and CSF biomarkers. This task is a three-class classification task distinguishing between mild cognitive impairment (MCI), control, and Alzheimer's disease (AD) patients. The MIMIC-IV (Johnson et al., 2023) and MIMIC Chest X-ray (MIMIC-CXR) (Johnson et al., 2019) datasets cover ICU visits and chest radiographs. We extracted clinical time-series data, chest X-ray images, demographics, and discharge notes, resulting in four modalities for a 25-class multi-label phenotyping task. The eICU collaborative database (Pollard et al., 2018) includes data from ICUs across the U.S. To predict morality, we focus on six tabular modalities: patient, diagnosis, treatment, medication, lab, and apacheApsVar (Acute Physiology and Chronic Health Evaluation (APACHE) Acute Physiology Score (APS)) tables, resulting in a two-class classification task. For more details on dataset pre-processing, see Appendix C.

5. Results

Using three biomedical datasets, diverse in modalities, feature space, and classification tasks, we demonstrate that our method consistently reduces computational costs compared to self-attention and cross-attention while enhancing performance. This is shown on a four-modality and two six-modality datasets.

We used a pre-trained ClinicalBERT model for the text modality and appropriate neural network (NN) architectures for other modalities (CNN for images, LSTM for time series, and a multi-layer perceptron for tabular data) as the unimodal baseline architectures and as the unimodal encoders before multimodal integration. We performed significance testing between OvO attention and the next best-performing model, detailed in Appendix E. For all models, we report the average of 10 random seeds with the respective metrics reported by the studies solving each task. FLOPs (*), were measured per sample and reported as the difference between concatenation and multimodal attention.

Results on MIMIC (Table 2) demonstrate OvO attention's scalability and performance advantages. OvO's 4.2M FLOPs significantly reduce computational costs compared to cross-attention (52.7M) and self-attention (67.6M), achieving reductions of 91.98% and 93.75%, respectively. Unimodal results show that the textual modality is most valuable in phenotype prediction, with ClinicalBERT alone outperforming self-attention and cross-attention. However, OvO attention extracts information from other modalities for a significant performance increase ($p < 0.01$).

Table 2. **MIMIC IV+CXR results**. Modalities: Time Series (TS), Image (I), Demographics (D), and Text (T). We use AUROC and AUPRC following (Hayat et al., 2022).

Model	Modalities	↓ Δ FLOPs	↑ AUROC	↑ AUPRC
LSTM	TS	-	58.8 ±0.6	28.5 ±0.4
CNN	I	-	56.9 ±0.3	26.7 ±0.2
NN	D	-	64.1 ±0.4	32.4 ±0.3
ClinicalBERT	T	-	79.3 ±0.4	58.7 ±0.3
Concatenation	All	*	82.7 ±0.6	65.1 ±1.8
Cross-Attention	All	52,723,712	78.2 ±2.1	54.1 ±2.7
Self-Attention	All	67,633,152	78.5 ±2.0	55.7 ±3.1
OvO Attention	**All**	**4,227,072**	**83.6 ±1.1**	**66.2 ±2.6**

For the six-modality Alzheimer's detection task (Table 3), OvO's 406M FLOPs significantly reduce computational costs compared to cross-attention (8.9M) and self-attention (9.6M), achieving reductions of 95.45% and 95.79%, respectively. Unimodal results show the cognitive tests modality is most valuable in disease prediction. OvO attention extracts information from other modalities for a significant performance increase ($p < 0.01$).

Lastly, results on the six-modality eICU mortality prediction task (Table 4) demonstrate OvO attention's scalability and performance advantages. OvO's 6.3M FLOPs significantly reduce computational costs compared to cross-attention (130M FLOPs) and self-attention (152M FLOPs), achieving reductions of 95.12% and 95.82%, respectively. Similar to the MIMIC and TADPOLE results, the dominant unimodal modality is Lab. OvO attention reflects these performance gains significantly ($p < 0.01$).

Table 3. **TADPOLE results**. Modalities: AV45 PET ROIs (A), CSF biomarkers (CSF), MRI ROIs (M), FDG PET ROI (F), Demographics (D), and Cognitive Tests (CT). We use multi-class area under the receiver operating curve (mAUC) and balanced classification accuracy (BCA), following (Marinescu et al., 2019).

Model	Modalities	↓ Δ FLOPs	↑ MAUC	↑ BCA
NN	A	-	63.5 ±3.1	56.4 ±3.8
NN	CSF	-	64.4 ±1.1	53.6 ±2.7
NN	M	-	67.0 ±1.3	57.2 ±1.0
NN	F	-	66.6 ±0.3	60.8 ±0.7
NN	D	-	74.6 ±0.9	62.0 ±0.6
NN	CT	-	97.8 ±0.2	88.6 ±0.7
Concatenation	All	*	97.7 ±0.8	91.9 ±1.9
Cross-Attention	All	8,921,088	97.1 ±0.6	90.7 ±1.7
Self-Attention	All	9,633,792	94.8 ±1.1	86.6 ±2.6
OvO Attention	**All**	**405,504**	**98.3 ±0.4**	**93.0 ±1.4**

Table 4. **eICU results**. Modalities: apacheApsVar (A), Demographics (DE), Medication (M), Lab (L), Diagnosis (D), and Treatment (T). We use AUROC and AUPRC following (Sheikhalishahi et al., 2020).

Model	Modalities	↓ Δ FLOPs	↑ AUROC	↑ AUPRC
NN	DE	-	50.2 ±0.6	91.8 ±0.2
NN	M	-	56.3 ±1.3	93.1 ±0.3
NN	D	-	58.2 ±2.1	93.3 ±0.4
NN	T	-	66.1 ±0.5	94.8 ±0.1
NN	A	-	77.6 ±0.2	97.0 ±0.1
NN	L	-	81.5 ±0.4	97.0 ±0.1
Concatenation	All	*	81.7 ±1.6	97.5 ±0.3
Cross-Attention	All	129,957,888	77.6 ±1.6	95.4 ±0.3
Self-Attention	All	151,781,376	80.2 ±2.0	96.8 ±0.4
OvO Attention	**All**	**6,340,608**	**82.5 ±0.9**	**97.8 ±0.2**

5.1. Generalizability

While our work focused on the clinical domain, where OvO will be most impactful and relevant, our method is a domain-agnostic approach and can be highly effective in other multimodal scenarios. We use the Amazon reviews dataset (Ni et al., 2019), to demonstrate OvO's performance on a non-clinical dataset. The Amazon Reviews dataset includes review images, review text, and product metadata (e.g., price, category), with the goal of review sentiment classification (positive or negative). See Appendix **??** for more preprocessing details. The Amazon Reviews results are shown in Table 5 and demonstrate both the scalability and performance advantages of OvO attention. OvO's 523,520 FLOPs significantly undercut those of cross-attention (1,903,616 FLOPs) and self-attention (2,685,440 FLOPs), achieving reductions of **72.50%** and **80.51%**, respectively, thereby highlighting OvO's efficiency on the integration. Even though the efficiency gains scale with the number of modalities (see Figure 2), it is impressive to see significant improvements can be achieved for a dataset with just three modalities. Since the textual modality is most valuable in sentiment prediction, the performance of BERT alone is higher than concatenation, self-attention, and cross-attention. This indicates that the noise from metadata and images interferes with model performance. However, OvO attention can extract information from the other two modalities for a significant performance increase rather than a decrease (p-value $<$0.01).

Table 5. **Amazon Reviews results**. Modalities are Image (I), Text (T), and Tabular (Tb). We report the average of 10 random seeds for accuracy, F1-scores, and standard deviations. (*) FLOPs were measured per sample and reported as the difference between concatenation and multimodal attention. We offer improved performance across all metrics and reduce FLOPs by at least 72.50% compared to self and cross-attention.

Model	Modalities	↓ Δ FLOPs	↑ Accuracy	↑ F1-Score
Neural Net	Tabular Metadata	-	57.6 ±0.7	57.5 ±0.8
ResNet	Images	-	66.3 ±0.7	66.6 ±0.6
BERT	Text	-	92.6 ±0.5	92.9 ±0.4
Concatenation	All	*	92.2 ±0.4	92.8 ±0.3
Self-Attention	All	2,685,440	92.4 ±0.4	92.4 ±0.4
Cross-Attention	All	1,903,616	91.6 ±0.7	92.2 ±0.6
OvO Attention	**All**	**523,520**	**93.1 ±0.3**	**93.0 ±0.3**

Overall, the adaptability and performance enhancement demonstrated by OvO across various applications, not limited to the medical field, highlight its potential as a universal scalable solution in multimodal learning. OvO's domain-agnostic approach can significantly contribute to overcoming the computational bottlenecks commonly encountered in multimodal scenarios.

6. Conclusion

We present One-Versus-Others (OvO), a new scalable multimodal attention mechanism. The proposed formulation significantly reduces the computational complexity compared to the widely used early fusion through self-attention and cross-attention methods. Notably, OvO achieves, at minimum, a reduction of 91.98% in FLOPs when benchmarked against self and cross-attention methods across a range of biomedical datasets encompassing up to six modalities. We provide both a detailed theoretical complexity analysis and empirical evidence from a simulated experiment, illustrating that OvO's computational demand scales linearly with the number of modalities, in contrast to the quadratic scaling observed in other methods. Overall, the results establish that OvO not only significantly reduces computational expenses but also exceeds the performance of existing state-of-the-art fusion methodologies.

References

Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., and Gong, B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34: 24206–24221, 2021.

Braman, N., Gordon, J. W., Goossens, E. T., Willis, C., Stumpe, M. C., and Venkataraman, J. Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pp. 667–677. Springer, 2021.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Golovanevsky, M., Eickhoff, C., and Singh, R. Multimodal attention-based deep learning for alzheimer's disease diagnosis. *Journal of the American Medical Informatics Association*, 29(12):2014–2022, 2022.

Hayat, N., Geras, K. J., and Shamout, F. E. Medfuse: Multimodal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pp. 479–503. PMLR, 2022.

Ilyin, S. E., Belkowski, S. M., and Plata-Salamán, C. R. Biomarker discovery and validation: technologies and integrative approaches. *Trends in biotechnology*, 22(8): 411–416, 2004.

Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., and Horng, S. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 2019.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

LI, B., Nakaguchi, T., Yoshimura, Y., and Xuan, P. Robust multi-modal prostate cancer classification via feature disentanglement and dual attention. *Transactions of Japanese Society for Medical and Biological Engineering*, (Abstract):308–308, 2021.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Marinescu, R. V., Oxtoby, N. P., Young, A. L., Bron, E. E., Toga, A. W., Weiner, M. W., Barkhof, F., Fox, N. C., Golland, P., Klein, S., et al. Tadpole challenge: Accurate alzheimer's disease prediction through crowdsourced forecasting of future data. In *Predictive Intelligence in Medicine: Second International Workshop, PRIME 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 2*, pp. 1–10. Springer, 2019.

Ming, Y., Dong, X., Zhao, J., Chen, Z., Wang, H., and Wu, N. Deep learning-based multimodal image analysis for cervical cancer detection. *Methods*, 205:46–52, 2022.

Mohammed, M., Mwambi, H., Mboya, I. B., Elbashir, M. K., and Omolo, B. A stacking ensemble deep learning approach to cancer type classification based on tcga data. *Scientific reports*, 11(1):1–22, 2021.

Ni, J., Li, J., and McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack, C. R., Jagust, W. J., Shaw, L. M., Toga, A. W., et al. Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209, 2010.

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., and Hussain, A. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25, 2018.

Rahman, T., Yang, M., and Sigal, L. Tribert: Full-body human-centric audio-visual representation learning for visual sound separation. *arXiv preprint arXiv:2110.13412*, 2021.

Schulz, S., Woerl, A.-C., Jungmann, F., Glasner, C., Stenzel, P., Strobl, S., Fernandez, A., Wagner, D.-C., Haferkamp, A., Mildenberger, P., et al. Multimodal deep learning for prognosis prediction in renal cancer. *Frontiers in oncology*, 11:788740, 2021.

Seo, P. H., Nagrani, A., Arnab, A., and Schmid, C. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17959–17968, 2022.

Sheikhalishahi, S., Balaraman, V., and Osmani, V. Benchmarking machine learning models on multi-centre eicu critical care dataset. *Plos one*, 15(7):e0235424, 2020.

Silva, L. A. V. and Rohr, K. Pan-cancer prognosis prediction using multimodal deep learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 568–571. IEEE, 2020.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, pp. 6558. NIH Public Access, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yu, J., Li, J., Yu, Z., and Huang, Q. Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, 30(12):4467–4480, 2019.

Zhu, L. and Yang, Y. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8746–8755, 2020.

A. Computational Complexity Analysis for Multimodal Integration Schemes

In this section, we present the step-by-step details of the computational complexity analysis presented in Section 3.2. The analysis is done with respect to the size of the input modalities associated with the three paradigms used in our experimental setting: early fusion followed by self-attention, cross-modal attention, and One-Versus-Others (OvO) Attention.

A.1. Early Fusion

The early fusion approach involves first combining the modalities and then processing the concatenated sequence with the self-attention mechanism.

Step 1: Concatenation of Modalities.

Let k be the number of modalities and n be the feature-length of each modality.

$$\text{Total length after concatenation} = k \times n$$

The complexity for this operation is linear:

$$\mathcal{O}(k \cdot n)$$

Step 2: Compute Queries, Keys, and Values.

The self-attention mechanism derives queries (Q), keys (K), and values (V) for the concatenated sequence (length $k \cdot n$) using linear transformations with representation dimension, d. The complexity of each transformation operation is:

$$\mathcal{O}(k \cdot n \cdot d)$$

Step 3: Compute Attention Scores.

Attention scores are computed by taking the dot product of queries and keys. The self-attention mechanism has quadratic complexity with respect to the sequence length and linear complexity with respect to the representation dimension d (Vaswani et al., 2017). Thus, given the concatenated sequence's length of $k \cdot n$ and the dimension of the keys and queries d, the complexity of this step is:

$$\mathcal{O}((k \cdot n)^2 \cdot d) = \mathcal{O}(k^2 \cdot n^2 \cdot d)$$

Step 4: Calculate the Weighted Sum for Outputs.

For each of the $k \cdot n$ positions in the concatenated sequence, we compute the softmax of the attention scores to produce the attention weights. These weights are then multiplied with their corresponding d-dimensional values to compute the weighted sum, which becomes the output. The computational complexity of these operations is:

$$\mathcal{O}(k^2 \cdot n^2 \cdot d)$$

When combining all steps, the dominating terms in the computational complexity stem from the attention scores' computation and the weighted sum, culminating in an overall complexity of:

$$\mathcal{O}(k^2 \cdot n^2 \cdot d)$$

A.2. Cross-modal Attention

For cross-modal attention, each modality attends to every other modality.

Step 1: Compute Queries, Keys, and Values for Inter-Modal Attention.

From a given modality, compute a query (Q), and from the remaining $k-1$ modalities, compute keys (K) and values (V). Keys, queries, and values are obtained using linear transformations with representation dimension d. The complexity of each transformation operation is:

$$\mathcal{O}(n \cdot d) \text{ for each query, key, value set}$$

Considering all modalities:

$$\mathcal{O}(k \cdot (k-1) \cdot n \cdot d)$$

The term $k \cdot (k-1)$ comes from the number of pairwise permutations of k, given by ${}_kP_2 = \frac{k!}{(k-2)!} = k(k-1)$.

Step 2: Calculate Attention Scores for Inter-Modal Attention.

The queries and keys from different modalities are used to compute attention scores, which represent how much one modality should attend to another.

$\mathcal{O}(n^2 \cdot d)$ for each pair of modalities (Vaswani et al., 2017)

Considering all modalities:

$$\mathcal{O}(k \cdot (k-1) \cdot n^2 \cdot d)$$

Step 3: Calculate the Weighted Sum for Outputs.

For every modality interaction, calculate the softmax of the attention scores to obtain the attention weights. These weights are then used in conjunction with the values vector to derive the weighted sum for the output:

$$\mathcal{O}(n^2 \cdot d) \text{ for each pair of modalities}$$

Considering all modalities:

$$\mathcal{O}(k \cdot (k-1) \cdot n^2 \cdot d)$$

When evaluating all steps together, the dominating factors in computational complexity arise from the computation of attention scores and the weighted sum. Thus, the collective complexity for cross-modal attention, where each modality attends to every other, equates to:

$$\mathcal{O}(k \cdot (k-1) \cdot n^2 \cdot d) = \mathcal{O}((k^2 - k) \cdot n^2 \cdot d)$$

For the complexity of cross-modal attention, the dominant term is k^2. The $k-1$ term effectively becomes a constant factor in relation to k^2. As k tends toward larger values, the difference between k^2 and $k^2 - k$ diminishes. This is a consequence of the principles of big O notation, which focuses on the fastest-growing term in the equation while dismissing constant factors and lower-order terms. As a result, for asymptotic analysis, the complexity

$$\mathcal{O}(k^2 - k) \cdot n^2 \cdot d$$

can be simplified to:

$$\mathcal{O}(k^2 \cdot n^2 \cdot d)$$

.

A.3. One-Versus-Others (OvO) Attention Complexity

Step 1: Averaging of ”Other” Modalities.
Let k be the number of modalities and n be the feature-length of each modality. For each modality m_i, averaging over the other $k-1$ modalities results in a complexity of:

$$\mathcal{O}(n)$$

Given that this needs to be computed for all k modalities:

$$\mathcal{O}(k \cdot n)$$

Step 2: Calculate Attention Scores with Shared Weight Matrix W.
The modality vector m_i and the average of ”other” modalities, $\frac{\sum_{j \neq i}^{n} m_j}{n-1}$, are used to compute attention scores, which represent how much one modality should attend to the others. Multiplication with the weight matrix W (with representation dimension d) and the dot product with the summed modalities lead to:

$$\mathcal{O}(n^2 \cdot d)$$

Considering this operation for all k modalities:

$$\mathcal{O}(k \cdot n^2 \cdot d)$$

Step 3: Calculate the Weighted Sum for Outputs.
For every modality interaction, calculate the softmax of the attention scores to obtain the attention weights. These weights are then used in conjunction with the m_i vector (analogous the values (V) vector) to derive the weighted sum for the output:

$$\mathcal{O}(n^2 \cdot d) \text{ for each pair of modalities}$$

Considering all modalities:

$$\mathcal{O}(k \cdot n^2 \cdot d)$$

When evaluating all steps together, the dominating factors in computational complexity arise from the computation of attention scores. Thus, the collective complexity for cross-modal attention, where each modality attends to every other, equates to:

$$\mathcal{O}(k \cdot n^2 \cdot d)$$

In summary, One-Versus-Others (OvO) Attention exhibits a computational complexity that grows linearly with respect to the number of modalities ($\mathcal{O}(k \cdot n^2 \cdot d)$). In contrast, both early fusion through self-attention and cross-attention approaches demonstrate quadratic growth with respect to the number of modalities ($\mathcal{O}(k^2 \cdot n^2 \cdot d)$). This makes OvO a more scalable option for multimodal integration.

A.4. Multi-headed OvO Attention

We extend OvO attention to the multi-headed attention framework to directly compare with early fusion through self-attention and pairwise cross-attention. Multi-headed attention allows the model to attend to the input embeddings in different ways simultaneously. This is achieved by splitting the input embeddings into multiple linear projections, each processed independently through a self-attention mechanism. The outputs of each attention head are then combined to obtain the final output of the multi-headed attention layer. Formally, taking the input modality m_i with respect to a set of other modalities ($m_j : j \neq i$), the multi-headed attention layer for OvO attention is defined as follows:

$$\left\{ \begin{array}{l} \mathit{MultiheadedOvO}(m_i, \{m_j : j \neq i\}) \\ = \mathit{concat}(\mathit{head}_1, \ldots, \mathit{head}_h) W^O \\ \mathit{head}_k = \mathit{OvO}(m_i W_k^{m_i}, \{m_j W_k^{m_j} : j \neq i\}) \end{array} \right. \tag{5}$$

Here, h is the number of attention heads, W_k is a learnable weight matrix for the k-th attention head, W^O is a learnable weight matrix that projects the concatenated outputs of the attention heads back to the original dimension, and OvO Attention is defined in Equation 4.

B. Simulation Dataset Details

We consider two classes: (1) 20 random feature values that sum up to 1.0, and, (2) 20 random feature values that are each less than 0.15. The threshold was chosen at 0.15 because if 0.10 was the threshold, the mean of the 20 values would be 0.05, and thus, the sum would also be very close to 1, on average. This would render the task too difficult, and there would not be a significant difference between the samples across the two labels. Setting the threshold to 0.2 would render the task too easy, as on average, the numbers are consistently greater in the second class and the classes could be differentiated using only one modality. Thus, we chose 0.15 as the threshold.

C. Dataset Descriptions and Detailed Pre-processing

C.0.1. The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) dataset

The Alzheimer's Disease Neuroimaging Initiative (ADNI) (Petersen et al., 2010) database provides neuroimaging data, cognitive test scores, biomarker profiles, and genetic information for Alzheimer's disease (AD), mild cognitive impairment (MCI), and normal patients. We use the processed data from the Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge (Marinescu et al., 2019). We utilize six modalities that have the least missing information per patient: cognitive tests - neuropsychological tests administered by a clinical expert; MRI ROIs (generated from Freesurfer) - measures of brain structural integrity; FDG PET ROI averages - measure cell metabolism, where cells affected by AD show reduced metabolism; AV45 PET ROI averages - measures amyloid-beta load in the brain; demographic information (e.g., age, gender, education); and CSF biomarkers - amyloid and tau levels in the cerebrospinal fluid. The preprocessing provided by TADPOLE turned every modality into a tabular form (including imaging). After removing patients with missing modalities, we had 767 MCI patients, 493 normal patients, and 143 AD patients. Thus, this is a three-class classification task with six modalities.

C.0.2. MIMIC-IV and CXR

MIMIC-IV (Johnson et al., 2023) covers 431K visits for 180K patients admitted to the ICU in the Beth Israel Deaconess Medical Center. MIMIC Chest X-ray (MIMIC-CXR) (Johnson et al., 2019) contains chest radiographs in DICOM format with free-text radiology reports. The dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at Beth Israel Medical Center. We follow the pre-processing of MedFuse (Mohammed et al., 2021) to extract the clinical time-series data from MIMIC-IV along with the associated chest X-ray images in MIMIC-CXR. We further expand the number of modalities by adding a demographics table and discharge notes, resulting in four modalities. We also follow MedFuse in the construction of the phenotyping task. The goal of this multi-label classification task is to predict whether a set of 25 chronic, mixed, and acute care conditions are assigned to a patient in a given ICU stay. This is a 25-class multi-label task with four modalities.

C.1. eICU Modality Descriptions and Detailed Pre-processing

The patient table contains patient demographic, admission, and discharge details. We use this table to determine mortality status. The diagnosis table contains active diagnoses given to each patient. We extract diagnosis features by one-hot encoding conditions specified in the 'diagnosisstring' column. The treatment table contains active treatments prescribed for each patient. We extract treatment features by one-hot encoding treatment types specified in the 'treatmentstring' column. The medication contains active medication orders for patients and when they were ordered. We extract medication features by one-hot encoding drugs specified in the 'drugname' column. The lab table contains lab type and corresponding results for various lab measurements collected for each patient. We extract lab features by summing lab measurements of commonly recorded lab types for each patient stay.

D. Implementation Details

D.1. Hyperparameter Tuning

For each experiment, we used validation accuracy to determine the best hyperparameters. We tuned the learning rate (0.01 - 1×10^{-8}), batch size (16, 32, 64, 128), epochs (200 epochs with early stopping if validation accuracy did not increase for 5 epochs), and the number of attention heads for OvO, self-attention, and cross-attention models (1, 2, 4, 8, 16). For the neural network encoders, we tuned the number of linear layers (1 to 4), and for the convolutional neural network, we tuned the number of convolution layers (1 to 4).

Our hyperparameter tuning scheme was consistent across all datasets and models. We used evaluation metrics on the validation set to determine the best hyperparameters. For compute times and GPU details used for hyperparameter tuning, see Appendix D.2.

We randomly picked 10 random seeds for each experiment. Once the best hyperparameters were selected, we ran ten models initialized with those seeds and parameters. The test set evaluation was then performed using these ten trained models, and the average results along with standard deviations are reported in Section 5.

D.2. Compute Resources

For each experiment, we use one NVIDIA GeForce RTX 3090 GPU. For the MIMIC task, single-modality models ran for roughly 40 minutes, and multi-modal models ran for roughly 55 minutes on average. For the eICU, the single modality pre-trained models ran for roughly 50 minutes, the single modality neural network ran for a minute, and the multi-modal models ran for approximately an hour on average. For the TADPOLE task, single-modality models ran for 5 minutes, while multi-modal models ran for roughly 15 minutes on average. In the simulation dataset, the maximum modalities was 20 which took our model, OvO, roughly 2 minutes to run, while the cross-modal attention baseline took about 20 minutes to run on average.

Task	Models	Runtime (minutes)
MIMIC	Unimodal	40
	Multimodal	55
eICU	Unimodal pre-trained	50
	Unimodal neural net	1
	Multimodal	60
TADPOLE	Unimodal	5
	Multimodal	15
Simulation	OvO (20 modalities)	2
	Cross and Self-attention	20

Table 6. Average runtimes for different tasks and model types using one NVIDIA GeForce RTX 3090 GPU.

D.3. Baselines

Our multimodal baselines include a conventional concatenation fusion with no attention, early fusion followed by self-attention, and pairwise cross-attention fusion. The architectures of all models are identical except for their integration stage. For example, since modality-specific encoders can produce different dimension sizes, we add a linear layer before integration to create the same input dimensions. While this step is not strictly necessary for concatenation, we still add the layer there so that no additional factors influence computation costs and performance. While there are many multimodal Transformers available for the vision-language domain, our focus is on examining the underlying fusion mechanism and creating a general integration paradigm for any application, especially ones outside of vision-language.

D.4. Data Splits and Evaluation Metrics

For the MIMIC dataset, we follow the established train, validation, and test split in (Hayat et al., 2022). Similarly, for the TADPOLE task, we use the provided data splits, but add a constraint that repeating patients cannot appear across data splits, as to avoid information leakage. In the other datasets, for consistency, we randomly sampled 80% of the data for the training set and 10% each for test and validation sets, as there was not an established and publicly accepted split. To evaluate our model against other integration techniques, we use the domain-accepted metrics for each task: For MIMIC and eICU, we use area under the receiver operating characteristic (AUROC) and area under Precision-Recall (PR) curve (AUPRC) as established in past works (Hayat et al., 2022; Sheikhalishahi et al., 2020); For TADPOLE we use the multi-class area under the receiver operating curve (mAUC) and the overall balanced classification accuracy (BCA), as established by the competition creators (Marinescu et al., 2019). For all datasets, we used the number of floating-point operations (FLOPs) as the measure of runtime complexity. FLOPs were measured per sample and reported as the difference between concatenation, the simplest integration setting, and multimodal attention (ΔFLOPs).

E. Significance Testing

We use a t-test to determine if there is a significant difference the performance metrics (AUROC, AUPRC, MAUC, BCA) means between OvO attention and the next best-performing multimodal model. Our sample size is 10 from each group, as we initialized the models with 10 random seeds. For the MIMIC IV and CXR dataset, we compare against self-attention as it performed the second best after OvO. Using an $\alpha = 0.01$, we have evidence to reject the null hypothesis and conclude that there is a statistically significant difference in means between single-attention and OvO attention. The p-value for the AUROC scores is 0.00363 and the p-value for AUPRC is 0.000948. For the TADPOLE challenge, we compare against cross-attention as it performed the second best after OvO. We get a p-value for MAUC scores of $2.09e^{-7}$ and a p-value of $8.19e^{-12}$ for BCA. Thus, we demonstrate a statistically significant difference in MAUC and BCA means between self-attention and OvO attention. Lastly, for the eICU dataset, we compare against cross-attention as it performed the second best after OvO. We get a p-value for AUROC scores of $2.00e^{-15}$ and a p-value of $8.24e^{-14}$ for AUPRC. Thus, we demonstrate a statistically significant difference in AUROC and AUPRC means between self-attention and OvO attention.

Dataset	Comparison Model	Metric	p-value
MIMIC	Self-attention	AUROC	0.00363
		AUPRC	0.000948
TADPOLE	Cross-attention	MAUC	2×10^{-7}
		BCA	8×10^{-12}
eICU	Cross-attention	AUROC	2×10^{-15}
		AUPRC	8×10^{-14}

Table 7. Results of t-tests comparing the performance metrics between OvO attention and the next best-performing multimodal models across different datasets.