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

Michal Golovanevsky[§], Eva Schiller[§], Akira Nair[§], Eric Han[§], Ritambhara Singh[§‡] [†*]

[§]*Department of Computer Science, Brown University,*
[‡] *and Center for Computational Molecular Biology, Brown University*
Providence, RI 02912, USA
[†] *E-mail: ritambhara@brown.edu*

Carsten Eickhoff*

School of Medicine,
Institute for Bioinformatics and Medical Informatics,
University of Tübingen,
Tübingen, 72074, Germany
E-mail: c.eickhoff@acm.org
*Co-corresponding authors

Multimodal models have become increasingly important as they surpass single-modality approaches on diverse tasks ranging from question-answering to disease diagnosis. 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 healthcare domain, 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 state-of-the-art multimodal 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 clinical datasets with multiple diverse modalities, we show that our method decreases computation costs while maintaining or increasing performance compared to popular integration techniques. Across all clinical datasets, OvO reduced the number of required floating point operations (FLOPs) by at least 91.98%, demonstrating its significant impact on efficiency and enabling multi-modal predictions in healthcare.*

Keywords: Multimodal learning; deep learning; attention mechanism; clinical decision support.
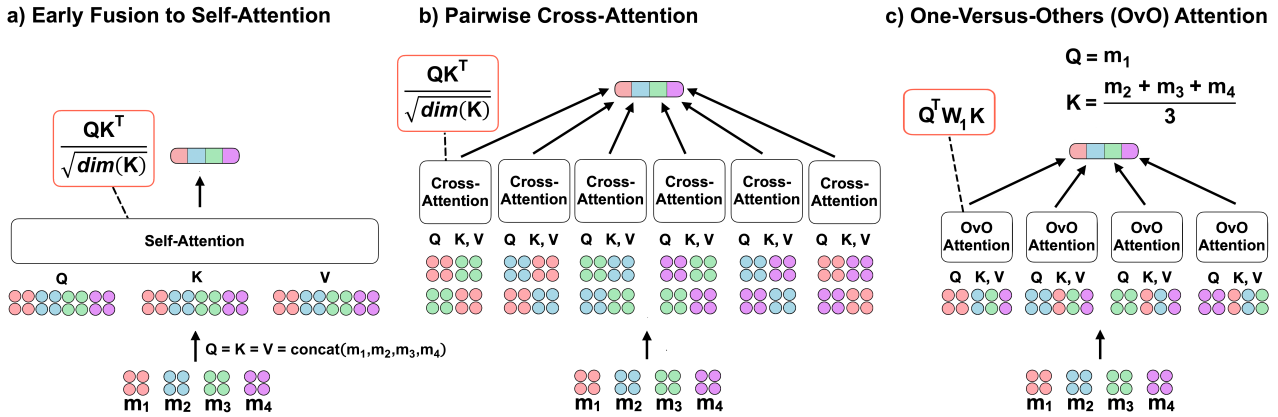
---

1. Introduction

Fig. 1. **Integration scheme comparison**. (a) Early fusion to self-attention with scaled dot product attention,[1] and (b) Pairwise cross-attention integration with scaled dot product attention.[1] (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).

Multimodal learning has emerged as a promising approach, which enables joint learning from multiple modalities of data (e.g., text and images). Combining different modalities allows for a more comprehensive and accurate understanding of tasks such as clinical decision support,[2–4] image and video captioning,[5,6] audio-visual speech recognition,[7] and sentiment analysis.[8] Multimodal learning has been explored through various methods in machine learning and deep learning. While feature-level integration was mostly used in more traditional machine learning algorithms, Neural Networks have allowed for the intermediate fusion of modalities at any layer and late fusion at the decision-making stage. However, these fusion paradigms lack a key component - capturing explicit interaction between modalities. For example, in detecting Alzheimer's Disease, genetic features help reinforce and ground the clinical information and thus lead to more robust decision-making.[3] Such relevant interactions can be captured through the attention mechanism. Popular multimodal models, such as LXMERT[9] and BLIP,[10] use a fusion method that captures interactions between modalities using *cross-attention*. On the other hand, models such as VisualBERT[11] and LLaVA[12] use early fusion, where vision and language inputs are concatenated early to learn multimodal through *self-attention*. The clinical domain embraced these approaches, with multimodal models like Med-FuseNet[13] and ARMOUR[14] using cross-attention for medical vision question answering and mortality prediction. In parallel, models such as BioViL-T[15] and MMBERT[16] employ early fusion through self-attention for disease prediction and report generation.

However, both self-attention and cross-attention grow quadratically in computational burden with the number of modalities, posing a scalability challenge. While in popular vision-language integration tasks, the number of modalities rarely exceeds four (images, text, audio, video), a significant bottleneck can arise in other domains. The healthcare domain exempli-

fies this issue, as a single task may involve integrating data from complex and rich sources spanning multiple modalities from radiology, pathology, genomics, genetics, and clinical data. Therefore, with the influx of many modalities, the use of cross-attention or self-attention will remain limited in the clinical domain as their computational demands escalate even further. To address this gap, we propose a new attention mechanism, One-Versus-Others (OvO) attention. OvO attention is calculated by comparing *one* modality against a combined representation of all *other* modalities (hence the name, One-Versus-Others). Our approach significantly reduces computational complexity as it grows linearly with the number of modalities (see Section 3.3). Figure 1 sketches a four-modality example to demonstrate the difference between our approach (scales linearly) and self-attention/cross-attention (scales quadratically). OvO is a general attention scheme that can be integrated into existing clinical multimodal architectures instead of cross-attention or self-attention.

We first, present a complexity analysis and validate it through a simulated dataset. Our simulation results show scalability gains in an extreme multimodal setting (20 modalities). Next, we use three diverse clinical datasets that vary in modalities, encoder types (pre-trained and not), number of samples, and tasks to show our model's improved scalability in different clinical multimodal settings. Our results demonstrate that our method dramatically decreases computation costs (offering at minimum a 91.98% reduction in computations), compared to self-attention and cross-attention while maintaining or even exceeding performance.

Overall, OvO is a novel attention scheme for multimodal integration that scales linearly to the number of modalities, enabling the practical application of deep learning models in healthcare, where computational efficiency and accuracy are vital for deployment.

2. Related work

Multimodal attention-based models are increasingly pivotal in clinical decision support systems, paralleling their widespread use in vision-language applications. In the medical domain, these models have shown remarkable utility in diverse scenarios, such as cancer classification,[17] biomarker discovery,[18,19] prognosis prediction,[20,21] and more. These applications highlight the versatility and potential of multimodal learning in handling complex and rich medical data. The attention mechanism serves as a core component in these models. Attention measures the similarity among individual representations, like word embedding vectors or, in the multimodal scenario, modality-specific embeddings. Each input embedding can assume one of three roles: (1) Query (Q), representing the current focus of attention when compared against other input embeddings; (2) Key (K), signifying an input embedding being compared to the Query; and (3) Value (V), which contributes to computing the output for the Query.

Commonly, the representations from each modality in the multimodal models are passed through one of two paradigms - early fusion followed by self-attention or fusion through cross-attention. The early fusion group (e.g., Transformer-based models like UNITER,[22] Visual-BERT,[11] LLaVA,[12] BioViL-T,[15] MedViLL,[23] etc.) concatenates the visual embeddings and the textual embeddings as a single input, before passing the inputs 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, Equation 1 shows the early fusion paradigm.

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

The cross-attention scheme (used in Transformer-based models like ViLBERT,[24] LXMERT,[9] MedFuseNet,[13] MADDi,[3] etc.) inputs each modality into its own Transformer, the outputs of which are fed to a cross-modal Transformer (see Figure 1 (b)). For such models, the cross-modal interactions are captured in a pairwise manner through cross-attention, where queries (Q), keys (K), and values (V) are computed from the modality inputs (m_1 and m_2), and then the keys and values from each modality are fed to the multi-headed attention block of the other modality. The output, Z, is shown in Equation 2.

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

While the early fusion and cross-attention paradigms could be extended to three modalities, seen in TriBERT[25] and VATT,[26] these models face scalability challenges for more than three modalities. Cross-attention methods can leverage joint representations formed from cross-attention but do not scale well to larger numbers of modalities as they are computed in a pairwise fashion. Thus, if there are k modalities, computing pairwise fusion between each pair will result in $\binom{k}{2}$ matrix computations. Moreover, attention is not a symmetric calculation, which means that most commonly, it is computed bi-directionally (e.g., image to text and text to image), leading to an even greater computational burden. Early fusion involves the concatenation of modalities before the Transformer layer, which similarly does not scale well with the number of modalities. Self-attention is quadratic with respect to sequence length,[1] and since early fusion methods concatenate inputs before attention, the computational complexity will increase quadratically as the number of modalities increases (see Section 3.3). Furthermore, concatenation is not order invariant, making the ordering of modalities an important consideration, potentially requiring similar bi-directional computations as cross-attention. Our integration method, OvO, addresses the limitations mentioned above 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 would be the case for cross-attention or self-attention (see Section 3.3). OvO computes attention between one modality at a time with respect to all other modalities. Given modality m_i, which is an embedding obtained from a dedicated encoder (e.g., CNN, ClinicalBERT, etc.) and $i \in 1, 2, \ldots, k$ where k is the number of modalities, OvO takes in one modality and computes the dot product against all the other modalities with a weight matrix W_i. W_i is a learnable parameter that can help scale the importance of each attention calculation (see Figure 1 (c)) and can learn

interactions throughout training. The modality embeddings m_i in OvO attention function like queries, while the weighted sum of the other modalities behaves like keys and values, akin to the dot-product attention mechanisms, with the naming choice of m_i adapted for multimodal applications. This dot product The similarity score function, representing the degree of alignment between the chosen modality and others, calculated for modality m_i with respect to a set of other modalities $(m_j : j \neq i)$ is shown in Equation 3. This produces a vector of scores which measure the relevance of m_i with respect to the other modalities. The context vector in OvO for modality m_i, which is a combined representation of information from the other modalities, is shown in Equation 4:

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

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

In Equation 4, the softmax is applied across the input dimension of the embeddings, normalizing across the attention scores. The result is then multiplied by the original modality embeddings to compute the final output. We chose to sum over the "other" modalities instead of concatenation because: (1) the concatenation vector will continue to increase in length with the number of modalities, which will result in a less scalable framework; (2) concatenation is not invariant to the order of modalities, which could affect the model prediction, whereas a sum provides position invariance.

Unlike cross-attention and self-attention, OvO provides a more interpretable mechanism for analyzing interactions between multiple modalities. In cross-attention, interactions are captured in isolated pairs (e.g., m_1 compared to m_2 or m_1 compared to m_3), limiting the ability to see how one modality integrates information from all others. Similarly, self-attention condenses modality interactions into a single operation, which can obscure explicit cross-modal interactions and make it difficult to disentangle their contributions. OvO, however, creates one attention matrix per modality, with each modality interacting with all others through the learnable weight matrix W_i, which ensures a flexible and adaptive scaling mechanism. In future work, W_i will be studied to better understand the relative importance of each modality's contributions to the final prediction.

3.2. *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:

$$\begin{cases} MultiheadedOvO(m_i, \{m_j : \ j \ \neq \ i\}) \\ = concat(head_1, \dots, head_h)W^O \\ head_k = OvO(m_iW_k^{m_i}, \{m_jW_k^{m_j} : \ j \ \neq \ i\}) \end{cases} \quad (5)$$

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 Equation 5 defines OvO Attention.

3.3. *Model Complexity*

This section highlights the complexities of the two commonly used paradigms: early fusion followed by self-attention and pairwise cross-attention, as well as our One-Versus-Others (OvO) attention. Table 1 summarizes the complexity per layer. Let k represent the number of modalities, n be the feature-length of each modality (assuming equal), and d be the representation dimension of the respective weight matrices. As established in,[1] self-attention has complexity of $\mathcal{O}(n^2 \cdot d)$. In the multimodal case, self-attention concatenates modalities before attention, leading to a sequence length of $k \cdot n$, influencing the quadratic term. Thus, the complexity of self-attention is $\mathcal{O}((k \cdot n)^2 \cdot d) = \mathcal{O}(k^2 \cdot n^2 \cdot d)$. Cross-attention computes attention over all pairwise permutations of modalities: $_kP_2 = \frac{k!}{(k-2)!} = k(k-1)$. Thus, the number of operations required by cross-attention is $\mathcal{O}(k \cdot (k-1) \cdot n^2 \cdot d) = \mathcal{O}((k^2 - k) \cdot n^2 \cdot d)$. When focusing on the fastest-growing terms in big O notation, the final complexity per layer simplifies to $\mathcal{O}(k^2 \cdot n^2 \cdot d)$. One-Versus-Others (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 1 provides step-by-step details for the complexity calculations.

Table 1. **Per-Layer complexities of model paradigm.** Let k be the number of modalities, n the feature-length of a modality, and d the representation dimension.

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

3.4. *Illustration through simulation*

To illustrate the linearity of OvO compared to the other integration paradigms, we simulated 20 artificial modalities. 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. These classes were created such that the correct label can only be inferred after inspecting all features. For example, 0.14

is less than 0.15, but it could also be a value that adds to 1. For more details on the simulation dataset and how the threshold was chosen, see Appendix Section 2.
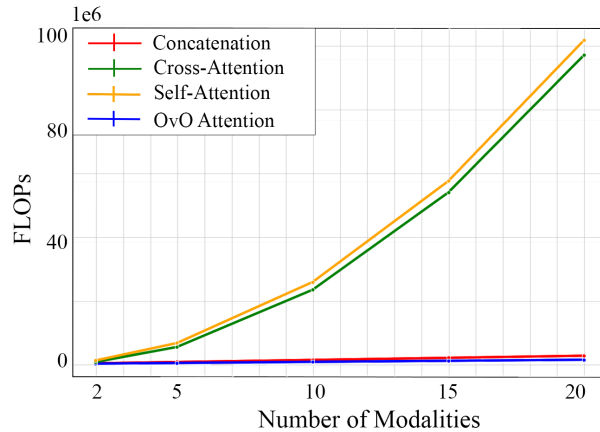

Fig. 2.   **The impact of using OvO attention to fuse simulated data.** Using FLOPs as a measure of compute, we demonstrate that OvO grows linearly with respect to the number of modalities, while self and cross-attention grow quadratically.

Each value was then vectorized by sampling randomly around the selected number, such that each modality is a vector of size 20 rather than a single number, leading to a combined total of 400 features. Overall, the dataset contains 2,000 samples (1,000 for each class). Our constructed simulation dataset tests the scaling capabilities of our method to an extent that real-world datasets do not usually reach.

We examine the computation cost across the three integration methods using 2, 5, 10, 15, and 20 simulated modalities. Most 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.

4. Experiments

We used three diverse clinical 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 clinical tasks feature a range of rich modalities that, despite their high integration costs, remain essential to solve.

4.1. *Dataset descriptions*

4.1.1. *MIMIC-IV and CXR data*

MIMIC-IV[27] covers 431K visits for 180K patients admitted to the ICU in the Beth Israel Deaconess Medical Center. MIMIC Chest X-ray (MIMIC-CXR)[28] 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[29] 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.

4.1.2. *The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) data*

The Alzheimer's Disease Neuroimaging Initiative (ADNI)[30] 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.[31] We focus on a one-time diagnosis prediction task, using the most recent available data point for each patient across all modalities. This ensures that each data sample includes information from the same time point, which aligns with our goal of evaluating cross-modal integration rather than longitudinal progression. 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.

4.1.3. *eICU data*

The eICU collaborative database includes data from ICUs across the continental United States between 2014 and 2015.[32] It consists of tables linked through a patient unit stay ID. For our mortality prediction task, we focus on six tabular modalities: patient, diagnosis, treatment, medication, lab, and apacheApsVar tables. The apacheApsVar table contains numerical variables used to calculate the Acute Physiology Score (APS), an established method within the Acute Physiology Age Chronic Health Evaluation (APACHE) system for summarizing patient's severity of illness on ICU admission and predicting outcomes. The patient table includes demographic, admission, and discharge details, and is used to determine mortality status. The diagnosis table lists active diagnoses for each patient, the treatment table includes active treatments, and the medication table contains active medication orders. We extract the features from these tables by one-hot encoding the relevant conditions, treatment types, and drug names, respectively. The lab table includes lab results, with features extracted by summing commonly recorded lab types. Our dataset includes 75,845 unique patients with 93,784 ICU stays, 86,012 recorded as alive and 7,772 as dead. This is a 2-class classification task with six modalities.

4.2. *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. Although 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. In Appendix Section 5, we touch on the limitations of our experiments and future work that we did not cover.

4.3. *Implementation details*

For the MIMIC dataset, we follow the established train, validation, and test split in Hayat *et al.*[4] Similarly, for the TADPOLE task, we use the provided data splits but add a constraint that repeating patients cannot appear across data splits 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 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;[4,33] 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.[31] 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).

4.4. *Hyperparameter Tuning*

Our hyperparameter tuning scheme was consistent for each dataset and each model. For each experiment, we used the evaluation metrics on the validation set to determine the best hyperparameters. We tuned the learning rate (0.01 - 1 x 10^{-8}, dividing by 10 for each interval), batch size (16, 32, 64, 128), epochs (200 epochs with early stopping if validation performance did not increase for 5 epochs), and 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 ranging from 1 to 4. Similarly, for the convolutional neural network, we tuned the number of convolution layers ranging from 1 to 4. For compute times and GPU details used for hyperparameter tuning, see Appendix Section 2. Lastly, we randomly picked 10 random seeds for every experiment - once the best hyperparameters were picked, ten models initialized with those seeds and parameters were run. Then, using the trained ten models, we evaluated on the test set and took the average of the 10 runs along with the standard deviation, which is reported in Section 5.

5. Results

Using three real-world clinical datasets, diverse in terms of the number of modalities, feature space, and classification tasks, we demonstrate that our method consistently and drastically reduces computational costs compared to early fusion and pairwise fusion while simultaneously maintaining or enhancing performance. This is demonstrated on one four-modality dataset and two six-modality datasets.

For the four-modality MIMIC task, we used pre-trained ClinicalBERT model for the text modality and fine-tuned it for the unimodal baseline and the multimodal task, separately, ensuring adaptation in each setting. For all other modalities, we used the appropriate neural network architecture (i.e., CNN for images, LSTM for time series, and a multi-layer perception for all tabular data). We perform significance testing between OvO attention and the next best-performing model, detailed in Appendix Section 3.

Table 2. **MIMIC IV+CXR results.**(*) 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 93.73% compared to self and cross-attention.

Model	Modalities	↓ Δ FLOPs	↑ AUROC	↑ AUPRC
LSTM	Time Series	-	58.8 ±0.6	28.5 ±0.4
CNN	Images	-	56.9 ±0.3	26.7 ±0.2
Neural Net	Demographics	-	64.1 ±0.4	32.4 ±0.3
ClinicalBERT	Text	-	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**

The results on MIMIC are presented in Table 2, clearly demonstrating the scalability and performance advantages of OvO attention. OvO's 4,227,072 FLOPs notably reduce computational costs compared to cross-attention (52,723,712 FLOPs) and self-attention (67,633,152 FLOPs), achieving reductions by **91.98%** and **93.75%**, respectively, thus highlighting OvO's superior efficiency. The unimodal results show that the textual modality is most valuable in phenotype prediction, and ClinicalBERT alone performs better than self-attention and cross-attention. This indicates that the added complexity and forced interactions are not necessarily conducive to result quality. However, OvO attention can extract information from the other modalities for a significant performance increase rather than a decrease (p-value <0.01, see Appendix Section 3).

For the six-modality Alzheimer's detection task from TADPOLE, we show our results in Table 3. OvO's 405,504 FLOPs significantly undercut cross-attention (8,921,088 FLOPs) and

Table 3. **TADPOLE results**. (*) 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 95.45% compared to self and cross-attention.

Model	Modalities	↓ Δ FLOPs	↑ MAUC	↑ BCA
Neural Net	AV45 PET ROI	-	63.5 ±3.1	56.4 ±3.8
Neural Net	CSF Biomarkers	-	64.4 ±1.1	53.6 ±2.7
Neural Net	MRI ROIs	-	67.0 ±1.3	57.2 ±1.0
Neural Net	FDG PET ROI	-	66.6 ±0.3	60.8 ±0.7
Neural Net	Demographics	-	74.6 ±0.9	62.0 ±0.6
Neural Net	Cognitive Tests	-	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**

self-attention (9,633,792 FLOPs), achieving reductions of **95.45%** and **95.79%**, respectively, highlighting OvO's remarkable efficiency. Similarly to the MIMIC results, the unimodal results show that the cognitive tests modality is most valuable in disease prediction, and performs on its own better than self-attention and cross-attention. However, OvO attention can extract information from the other modalities for a significant performance increase rather than a decrease (p-value <0.01).

Lastly, the results on the six-modality eICU mortality prediction task are shown in Table 4, demonstrating the scalability and performance advantages of OvO attention.

OvO's 6,340,608 FLOPs significantly undercut those of cross-attention (129,957,888 FLOPs) and self-attention (151,781,376 FLOPs), achieving reductions of approximately **95.12%** and **95.82%**, respectively, thereby highlighting OvO's efficiency. Mirroring the trends observed in the MIMIC and TADPOLE datasets, we note a dominant unimodal modality, specifically Lab modality, in this experiment as well. While concatenating modalities does enhance performance, this improvement is not seen in self and cross-attention models. In contrast, OvO attention not only reflects these performance gains but does so significantly (p-value <0.01). We hypothesize that this is due to the overfitting of more complex integration frameworks of self and cross-attention on relatively smaller clinical datasets. OvO, in its simplicity akin to concatenation, manages to strike a balance by maintaining flexibility and capturing inter-modal interactions through its attention mechanism, thus offering an edge in performance without excessive complexity.

In summary, across diverse clinical datasets and modalities, OvO attention consistently outperforms traditional fusion techniques in both predictive performance and computational efficiency, underlining its robustness in handling complex multimodal healthcare data.

Table 4. **eICU results**. We report the average of 10 random seeds for AU-ROC and AUPRC, along with standard deviations. (*) FLOPs were measured per sample and reported as the difference between concatenation and multi-modal attention. We offer improved performance across all metrics and reduce FLOPs by at least 95.12% compared to self and cross-attention.

Model	Modalities	↓ Δ FLOPs	↑ AUROC	↑ AUPRC
Neural Net	Demographics	-	50.2 ±0.6	91.8 ±0.2
Neural Net	Medication	-	56.3 ±1.3	93.1 ±0.3
Neural Net	Diagnosis	-	58.2 ±2.1	93.3 ±0.4
Neural Net	Treatment	-	66.1 ±0.5	94.8 ±0.1
Neural Net	APACHE APS	-	77.6 ±0.2	97.0 ±0.1
Neural Net	Laboratory	-	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**

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 clinical datasets containing 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. Our proposed method provides a way to overcome one of the major challenges associated with multimodal datasets - computational resource demand and cost, thus enabling adoption in resource-constrained domains, such as clinical decision support. Overall, the results unequivocally establish that OvO not only significantly reduces computational expenses but also exceeds the performance of existing state-of-the-art fusion methodologies.

References

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* **30** (2017).
2. Y. Ming, X. Dong, J. Zhao, Z. Chen, H. Wang and N. Wu, Deep learning-based multimodal image analysis for cervical cancer detection, *Methods* **205**, 46 (2022).
3. M. Golovanevsky, C. Eickhoff and R. Singh, Multimodal attention-based deep learning for alzheimer's disease diagnosis, *Journal of the American Medical Informatics Association* **29**, 2014 (2022).

4. N. Hayat, K. J. Geras and F. E. Shamout, Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images, in *Machine Learning for Healthcare Conference*, 2022.
5. J. Yu, J. Li, Z. Yu and Q. Huang, Multimodal transformer with multi-view visual representation for image captioning, *IEEE transactions on circuits and systems for video technology* **30**, 4467 (2019).
6. P. H. Seo, A. Nagrani, A. Arnab and C. Schmid, End-to-end generative pretraining for multi-modal video captioning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
7. G. Sterpu, C. Saam and N. Harte, Attention-based audio-visual fusion for robust automatic speech recognition, in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018.
8. S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh and A. Hussain, Multimodal sentiment analysis: Addressing key issues and setting up the baselines, *IEEE Intelligent Systems* **33**, 17 (2018).
9. H. Tan and M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, *arXiv preprint arXiv:1908.07490* (2019).
10. J. Li, D. Li, C. Xiong and S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in *International conference on machine learning*, 2022.
11. L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh and K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, *arXiv preprint arXiv:1908.03557* (2019).
12. H. Liu, C. Li, Q. Wu and Y. J. Lee, Visual instruction tuning, *Advances in neural information processing systems* **36** (2024).
13. D. Sharma, S. Purushotham and C. K. Reddy, Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain, *Scientific Reports* **11**, p. 19826 (2021).
14. J. Liu, D. Capurro, A. Nguyen and K. Verspoor, Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities, *Journal of Biomedical Informatics* **145**, p. 104466 (2023).
15. S. Bannur, S. Hyland, Q. Liu, F. Perez-Garcia, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme *et al.*, Learning to exploit temporal structure for biomedical vision-language processing, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
16. Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar and C. Jawahar, Mmbert: Multi-modal bert pretraining for improved medical vqa, in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021.
17. B. LI, T. Nakaguchi, Y. Yoshimura and P. Xuan, Robust multi-modal prostate cancer classification via feature disentanglement and dual attention, *Transactions of Japanese Society for Medical and Biological Engineering* , 308 (2021).
18. N. Braman, J. W. Gordon, E. T. Goossens, C. Willis, M. C. Stumpe and J. Venkataraman, 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*, 2021.
19. S. E. Ilyin, S. M. Belkowski and C. R. Plata-Salamán, Biomarker discovery and validation: technologies and integrative approaches, *Trends in biotechnology* **22**, 411 (2004).
20. S. Schulz, A.-C. Woerl, F. Jungmann, C. Glasner, P. Stenzel, S. Strobl, A. Fernandez, D.-C. Wagner, A. Haferkamp, P. Mildenberger *et al.*, Multimodal deep learning for prognosis prediction in renal cancer, *Frontiers in oncology* **11**, p. 788740 (2021).

21. L. A. V. Silva and K. Rohr, Pan-cancer prognosis prediction using multimodal deep learning, in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020.
22. Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng and J. Liu, Uniter: Universal image-text representation learning, in *European conference on computer vision*, 2020.
23. J. H. Moon, H. Lee, W. Shin, Y.-H. Kim and E. Choi, Multi-modal understanding and generation for medical images and text via vision-language pre-training, *IEEE Journal of Biomedical and Health Informatics* **26**, 6070 (2022).
24. J. Lu, D. Batra, D. Parikh and S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, *Advances in neural information processing systems* **32** (2019).
25. T. Rahman, M. Yang and L. Sigal, Tribert: Full-body human-centric audio-visual representation learning for visual sound separation, *arXiv preprint arXiv:2110.13412* (2021).
26. H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui and B. Gong, Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, *Advances in Neural Information Processing Systems* **34**, 24206 (2021).
27. A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow *et al.*, Mimic-iv, a freely accessible electronic health record dataset, *Scientific data* **10**, p. 1 (2023).
28. A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz and S. Horng, Mimic-cxr-jpg-chest radiographs with structured labels, *PhysioNet* (2019).
29. M. Mohammed, H. Mwambi, I. B. Mboya, M. K. Elbashir and B. Omolo, A stacking ensemble deep learning approach to cancer type classification based on tcga data, *Scientific reports* **11**, 1 (2021).
30. R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga *et al.*, Alzheimer's disease neuroimaging initiative (adni): clinical characterization, *Neurology* **74**, 201 (2010).
31. R. V. Marinescu, N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, P. Golland, S. Klein *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*, 2019.
32. T. J. Pollard, A. E. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark and O. Badawi, The eicu collaborative research database, a freely available multi-center database for critical care research, *Scientific data* **5**, 1 (2018).
33. S. Sheikhalishahi, V. Balaraman and V. Osmani, Benchmarking machine learning models on multi-centre eicu critical care dataset, *Plos one* **15**, p. e0235424 (2020).