Orthogonal Sequential Fusion in Multimodal Learning

1st Sami Labbaki
Department of Computer Science
University of Oxford
Oxford, United Kingdom
sami.labbaki@cs.ox.ac.uk

2nd Peter Minary
Department of Computer Science
University of Oxford
Oxford, United Kingdom
peter.minary@cs.ox.ac.uk

***Abstract*—The integration of data from multiple modalities is a fundamental challenge in machine learning, encompassing applications from image captioning to text-to-image generation. Traditional fusion methods typically combine all inputs concurrently, which can lead to an uneven representation of the modalities and restricted control over their integration. In this paper, we introduce a new fusion paradigm called Orthogonal Sequential Fusion (OSF), which sequentially merges inputs and permits selective weighting of modalities. This stepwise process also enables the promotion of orthogonal representations, thereby extracting complementary information for each additional modality. We demonstrate the effectiveness of our approach across various applications, and show that OSF outperforms existing fusion techniques. Our approach represents a promising alternative to established fusion techniques and offers a sophisticated way of combining modalities for a wide range of applications, including integration into any complex multimodal model that relies on information fusion.**

***Index Terms*—Multimodal learning, Neural networks, Machine learning, Deep learning, Information theory, Data integration, Healthcare**

I. Introduction

Data integration across various modalities is a foundational pillar in machine learning and artificial intelligence. Central to this concept is the ambition to augment our understanding of complex and heterogeneous data, broadening its applications from straightforward unimodal tasks to more sophisticated endeavors such as the encoding and processing of comprehensive surrounding data in the context of human-computer interactions [1] or autonomous driving [2]. By concurrently processing and understanding different types of data, multimodal learning promises to yield more robust models that provide a comprehensive interpretation of data and result in more accurate predictions and insights [3], [4].

The ubiquity of big data in multimodal machine learning—spanning text, image, audio, video, and other modalities—has underscored the importance of effective fusion techniques [5]. However, realizing this potential is not without its challenges. Current methods, though numerous and highly advanced, often lack model-agnostic designs and can be computationally expensive, thereby complicating their integration into broader frameworks. Moreover, most fusion techniques process all inputs in a single step, leading to a rigid integration that may inadvertently emphasize or overshadow certain modalities. This concurrent processing can result in the dominance of a single modality, ultimately hindering the comprehensive capture of the diverse information that multimodal data can offer.

In response to these challenges, our work introduces Orthogonal Sequential Fusion (OSF), an innovative fusion paradigm aimed at overcoming the limitations of existing multimodal fusion techniques. Unlike traditional fusion methods, OSF operates in a sequential manner, integrating one modality at a time, thereby offering a flexible way to weight and prioritize individual modalities based on their relevance. This approach ensures a balanced representation and maximizes the extraction of complementary information across modalities via its custom loss function. OSF's stepwise process promotes orthogonal representations, effectively capturing the distinct and complementary information that each modality offers. This not only enhances the overall performance of the system but also provides valuable insights into the relationships and interactions between different modalities.

In this paper, we carry out an extensive series of experiments to illustrate the effectiveness of OSF in addressing a multitude of data integration scenarios. These experiments are carefully designed to encompass a broad spectrum of applications, including variations in the number of modalities and variation in model complexity. Moreover, we provide evidence of the practical effectiveness of OSF through its integration within a state-of-the-art model, highlighting its ability to enhance performance beyond traditional fusion techniques in a specific application. These experiments benchmark OSF against established fusion techniques, clearly illustrating that our approach outperforms these methods in terms of evaluation metric.

The contributions of this paper can be summarized as follow:

1) We introduce OSF, a novel fusion paradigm for multimodal machine learning. This approach allows for the stepwise integration and discretionary ordering of modalities, promoting a balanced representation and optimizing the extraction of complementary information from each modality.
2) We benchmark OSF against prevailing fusion approaches, on diverse datasets, and demonstrate that OSF consistently surpasses the baselines, exhibiting a partic-

ularly notable superiority within the context of a highly multimodal environment.

3) We incorporate OSF into a state-of-the-art model, illustrating its seamless integration even within highly sophisticated and complex models. In addition, our findings highlight the ability of this fusion paradigm to further boost performance in already high-achieving applications.

II. Related work

A. Multimodal Learning and Traditional Fusion Techniques

Traditional fusion methods have found widespread application in a variety of fields, spanning from computer vision and speech recognition to natural language processing and healthcare. These techniques, including early, late, and hybrid fusion methods, represent the foundational bedrock of multimodal fusion and have significantly influenced contemporary research. Early fusion, or feature-level fusion, merges the features of all modalities at the very beginning of the process [6]. Conversely, late fusion, often referred to as decision-level fusion, initially processes each modality separately, then combines these individual decisions to form a final conclusion [7]. Hybrid fusion, a balanced blend of early and late fusion, leverages the strengths of both methodologies, offering a more versatile approach to multimodal data integration [8]. Even in the face of increasingly complex models and methodologies, these traditional fusion techniques remain prevalently employed, often serving as the base upon which these more advanced models are built [9].

B. Limitations of Traditional Fusion Methods

While traditional fusion methods still hold a prominent place in multimodal learning, they are not without certain constraints. One of the limitations of these approaches is their proneness to process all modalities concurrently. This simultaneous processing can lead to an over-representation of dominant modalities and an under-representation of subtler ones, which can contain valuable information. These imbalances can weaken the overall efficacy and robustness of the fusion process, thereby contradicting the core aim of multimodal learning: to assimilate and harness the unique insights provided by complementarities and redundancies of the information conveyed by the modalities. Furthermore, traditional fusion techniques lack a mechanism for dynamic weighting or prioritization of individual modalities, a shortfall that becomes increasingly significant when dealing with data that is either imbalanced, noisy or highly multimodal. These identified limitations emphasize the need for fusion methodologies that offer a more adaptive and refined approach.

C. Recent Advancements in Multimodal Fusion

In recent years, efforts to address the shortcomings of conventional fusion techniques have encouraged the development of novel approaches to reshape the fusion process. A significant breakthrough in this area has been the introduction of attention mechanisms [10], enabling the fusion process to dynamically assess the importance of different modalities. This development allows the model to adapt to the unique attributes of each data instance, thereby tailoring its fusion strategy to optimize information extraction. Another prevalent focus revolves around creating shared or joint representations across modalities with the goal of establishing a more unified and integrated fusion process [11]. Nevertheless, these advancements, albeit significant, have not entirely resolved the task of achieving a balanced and comprehensive representation of several modalities. Moreover, these approaches often lack flexibility, being typically model-based [5], and may exhibit substantial computational resources and complexity in integration. Consequently, quantitatively benchmarking OSF against them is beyond the scope of this paper. However, these enduring challenges underline the ongoing need for further innovative thinking and exploration in the field, towards creating fusion techniques that are both efficient and flexible in dealing with multimodal data.

D. Orthogonality in Machine Learning

A promising direction for multimodal learning lies in the concept of orthogonality, a mathematical principle that has found diverse applications within machine learning research. Encouraging orthogonal representations can minimize redundancy and promote diversity within the feature space [12], [13], qualities that are especially desirable within the context of multimodal fusion. By ensuring that each modality contributes unique and non-redundant information to the fusion process, we can harness the full potential of multimodal data. Orthogonality has been employed to enhance various machine learning tasks, such as initialization [14] and representation learning [15]. However, its application in multimodal fusion, although very promising [16], is largely unexplored.

III. Methodology

A. Sequential Fusion

In multimodal fusion, the prevalent practice is to employ all modalities simultaneously for combining the extracted information from disparate sources. However, in this research work, we support the idea of sequentially fusing the modalities, step by step in a pairwise fashion. This sequential approach provides enhanced control and sophistication over the fusion process. Our inspiration for this process is drawn from the human reasoning mechanism, where when presented with three information sources, we can split the sources in groups to arrive at a well-informed decision [17].

Following the same logic, our method first fuses two modalities and then merges the resulting information with a third modality. This process is repeated until all modalities have been processed. Specifically, the entire fusion procedure follows a predefined modality order $\mu = (\mu_1, \mu_2, \ldots, \mu_N)$, where μ_t denotes the index of the modality fused at step t.

A pseudo code for pair-wise sequential fusion is given in Algorithm 1, and a high-level overview of the sequential fusion process is illustrated in Fig. 1.

By selectively combining modalities in a sequential manner, we can effectively exploit the underlying relationships between them and weigh their contribution to the final output accordingly by facilitating the extraction of complementarities and redundancies at each layer.

Algorithm 1 Sequential Fusion Algorithm for Pairwise Fusion

Input: N different ordered modalities $m_{\mu_1}, \ldots, m_{\mu_N}$
Output: A final embedding for all fused modalities

for $k = 1$ to N **do**
  $E_{\mu_k} \leftarrow \texttt{ENCODE}(m_{\mu_k})$ {embed the modalities in dimension d using dense layers}
end for
$x \leftarrow E_{\mu_1}$
for $k = 2$ to N **do**
  $y \leftarrow E_{\mu_k}$
  $x \leftarrow \texttt{FUSE}(x, y)$
end for
return x

B. Sequential Fusion as a Masking Framework

Bridging the gap between theoretical algorithms and practical implementations, OSF can be reformulated as a masking framework that selectively modulates the input from each modality, thereby controlling its influence on the final integrated representation.

Let $\mathbf{E} \in \mathbb{R}^{N \times d}$ denote the matrix of N modality embeddings, where each row $\mathbf{E}[t,:]$ is the embedding of the t-th modality, $E_{\mu_t} \in \mathbb{R}^d$. OSF operates by progressively unmasking and incorporating one modality at a time, while all other modalities remain masked. The fused representation is initialized as:

$$F_1 = E_{\mu_1}, \quad F_1 \in \mathbb{R}^d \tag{1}$$

Initially, all modalities except for E_{μ_1} are masked. At step t, OSF unmasks E_{μ_t} and fuses it with the current fused representation F_{t-1}. We represent this mask as a binary vector $\mathbf{M}_t \in \{0,1\}^N$, where

$$\mathbf{M}_t[i] = \begin{cases} 1, & \text{if } i = \mu_t, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Thus, the embedding E_{μ_t} used in the fusion process at step t is obtained by multiplying the row mask $\mathbf{M}_t$ by $\mathbf{E}$:

$$E_{\mu_t} = \mathbf{M}_t \, \mathbf{E} \in \mathbb{R}^d \tag{3}$$

For $t = 2, \ldots, N$, the fused representation F_t is updated by fusing F_{t-1} with the unmasked modality E_{μ_t}:

$$F_t = g_t\big(F_{t-1}, E_{\mu_t}\big) \tag{4}$$

where $g_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is a fusion function that combines the current fused representation with the newly unmasked modality at each step t. In our implementation, we use distinct dense layers, but more sophisticated fusion functions could also be used, as long as they adhere to the requirement $g_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$. Alternatively, a single shared function could be employed across all steps to reduce model complexity.

After all N modalities have been unmasked and fused, the final fused representation is:

$$F_N = g_N(F_{N-1}, E_{\mu_N}) \tag{5}$$

C. Modality ordering

Determining the optimal order for sequential fusion is a critical step in achieving effective fusion. While the most obvious approach is to test all the possible combinations for ordering the modalities, it is clear that this method can be computationally expensive and may not scale well as the number of modalities grows. In contrast, our ranking-based approach offers a simple and efficient solution for determining the fusion order while avoiding the computational burden of exhaustive search.

Our approach is based on training unimodal models, implemented here as simple feed-forward networks, and ranking the modalities based on their individual performances. This ranking provides an initial ordering of the modalities according to their relative strengths in the task at hand. We then fuse the modalities starting from the least performing and moving toward the best performing, aiming to capture complementary information from each modality and fuse it in a way that improves overall performance.

However, there are limitations to this approach. While unimodal training provides an indication of the individual performance of each modality, it may not fully capture the interactions and dependencies that exist between pairs of modalities. As a result, the ranking obtained by this approach may not always lead to the most optimal fusion order, and other methods such as exhaustive search may be necessary to explore all possible fusion orders. Moreover, our approach assumes that the modalities have distinct contributions to the task at hand and can be ranked based on their individual performances. In some cases, certain modalities may be highly correlated or redundant, and the ranking may not accurately reflect their true contributions to the task.

Despite these limitations, our ranking-based approach offers a practical and efficient solution for determining fusion order. In our experiments on the Carnegie Mellon University Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) dataset [18] with three modalities, we found that the ordering determined by our approach was the optimal among all possible orders.

D. Advanced Modality Ordering

While the methods discussed in the previous subsection is sufficient for simpler tasks or scenarios with fewer modalities, more advanced tasks and highly multimodal setups require a more sophisticated strategy. To address this, we propose another approach to modality ordering that uses comprehensive evaluations of unimodal and bimodal models, leveraging a metric that combines performance and stability measures.

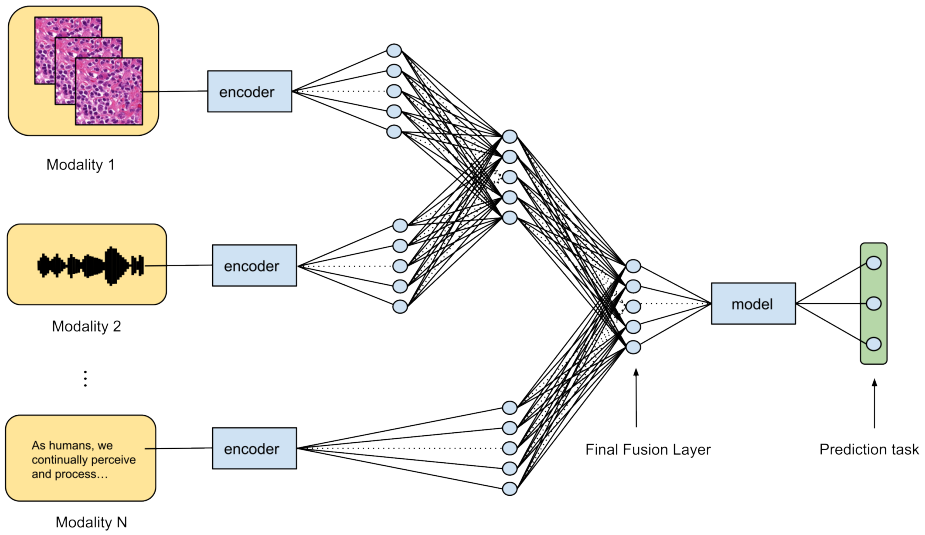

Fig. 1. Orthogonal Sequential Fusion with N modalities for a given prediction task.

Let $\mathcal{M} = \{m_1, m_2, \ldots, m_N\}$ denote a set of N modalities. The objective is to determine an optimal ordering $(m_{\mu_1}, m_{\mu_2}, \ldots, m_{\mu_N})$, where $m_{\mu_i} \in \mathcal{M}$.

For each modality $m_i \in \mathcal{M}$, a unimodal model is trained across S random seeds, producing performance metrics $c_{i,s}$. The aggregated metrics are defined as:

$$\tilde{c}_i = \text{median}(\{c_{i,s}\}_{s=1}^{S}), \quad \sigma_i^c = \text{std}(\{c_{i,s}\}_{s=1}^{S}) \tag{6}$$

The adjusted unimodal performance metric is:

$$\Delta_i = (\tilde{c}_i - \sigma_i^c) \tag{7}$$

Modality rankings are derived by sorting $\{\Delta_i\}_{i=1}^{N}$ in ascending order to produce the initial ordering:

$$\begin{aligned} &\mu_{\text{unimodal}} = (\mu_{u1}, \mu_{u2}, \ldots, \mu_{uN}) \\ &\text{where } \Delta_{\mu_{u1}} \leq \Delta_{\mu_{u2}} \leq \cdots \leq \Delta_{\mu_{uN}} \end{aligned} \tag{8}$$

For each pair of modalities (m_i, m_j) with $i \neq j$, bimodal models are trained across S random seeds. Aggregated metrics are computed as:

$$\tilde{c}_{ij} = \text{median}(\{c_{ij,s}\}_{s=1}^{S}), \quad \sigma_{ij}^c = \text{std}(\{c_{ij,s}\}_{s=1}^{S}) \tag{9}$$

Details on the computation of these aggregated metrics can be found in Subsection IV-E.

The adjusted pairwise performance metric is:

$$\Delta_{ij} = (\tilde{c}_{ij} - \sigma_{ij}^c) \tag{10}$$

Using $\{\Delta_{ij} \mid 1 \leq i < j \leq N\}$, a weighted graph $G = (\mathcal{M}, E, w)$ is constructed:

- **Nodes** $\mathcal{M}$: Represent modalities.
- **Edges** $E = \{(m_i, m_j) \mid i \neq j\}$: Represent modality pairs.
- **Weights** $w_{ij} = 1 - \Delta_{ij}$: Represent adjusted pairwise performance.

As the graph is complete, a Hamiltonian path necessarily exists. The top ten shortest Hamiltonian paths are identified based on edge weights w_{ij}. Among these candidate paths, we choose the one that maximizes agreement with μ_{unimodal}. Here, agreement is defined as the number of consecutive pairs (a, b) in the Hamiltonian path that also appear as adjacent elements (in either order) in the ordering derived from μ_{unimodal}. Since a Hamiltonian path does not impose a fixed direction, each path can be traversed in two possible orderings. From these, we select the optimal ordering by evaluating both orientations on the validation set.

The optimal path P_{optimal} defines the final modality ordering, which prioritizes both individual modality performance and synergistic interactions.

E. Orthogonal Loss Function

One of the major novelties of our proposed sequential fusion method lies in the loss function. While the sequential fusion process already allows us to combine modalities in groups and extract information at each step, we believe that it is even more beneficial to incentivize the modalities to provide complementary information. This is particularly relevant when some modalities encode overlapping information. To learn useful complementary features through the fusion process, we propose encouraging the latent representations to be orthogonal before fusion occurs.

OSF is inspired by L1-L2 regularization [19] and can be applied to any loss function. The objective of OSF is to encourage the fused embeddings at each layer to be orthogonal, which in turn helps to capture more diverse and complementary information from the modalities. To achieve this, we introduce hyperparameters, $\lambda_{1,t}$ and $\lambda_{2,t}$, which control the strength of the L1 and L2 regularization terms, respectively.

Assuming we have N modalities and $N-1$ fusion layers, where each layer L_t consists of the embeddings $[F_t, E_{\mu_{t+1}}]$, the loss function for OSF can be written as:

$$\begin{aligned}\mathcal{L}_{\text{OSF}} = \mathcal{L}(\text{preds}, \text{outputs}, T) \\ + \sum_{t=1}^{N-1} (\lambda_{1,t} \left| F_t \cdot E_{\mu_{t+1}} \right| \\ + \lambda_{2,t} (F_t \cdot E_{\mu_{t+1}})^2)\end{aligned} \tag{11}$$

where $\mathcal{L}$ represents the primary loss function for a given task T, and *preds* and *outputs* denote the model predictions and the ground truth, respectively.

The OSF loss function can be optimized using standard backpropagation algorithms. During training, the network learns to weigh the modalities in a way that maximizes the diversity and complementarity of the fused embeddings at each layer, resulting in a more effective and robust fusion process.

Overall, the use of the OSF loss function in our sequential fusion approach further improves the extraction process at each sequential step. By encouraging orthogonal embeddings, we can capture more diverse and complementary information from the modalities, leading to improved performance and more effective models.

F. Theoretical Foundations of OSF

Let $\{E_{\mu_1}, E_{\mu_2}, \ldots, E_{\mu_N}\}$ be a sequence of ordered embeddings, and let F_t denote the fused representation at step t. The Partial Information Decomposition (PID) of the information $I(F_t)$ is given by:

$$\begin{aligned}I(F_t) = I(F_{t-1}) + I(E_{\mu_t} \mid F_{t-1}) \\ - R(F_{t-1}, E_{\mu_t}) + S(F_{t-1}, E_{\mu_t})\end{aligned} \tag{12}$$

where:

- $I(F_{t-1})$ is the *existing* information in F_{t-1},
- $I(E_{\mu_t} \mid F_{t-1})$ is the *additional* information from E_{μ_t} given F_{t-1},
- $R(F_{t-1}, E_{\mu_t})$ is the *redundant* information between F_{t-1} and E_{μ_t},
- $S(F_{t-1}, E_{\mu_t})$ is the *synergistic* information emerging from their interaction.

This decomposition is more granular than the standard chain rule of Shannon information and is useful for analyzing how multiple representations combine or overlap [20], [21].

Inspired by deep learning decorrelation and orthogonalization methods [13], [22], we approximate redundancy as:

$$R(F_{t-1}, E_{\mu_t}) \propto |F_{t-1} \cdot E_{\mu_t}| + (F_{t-1} \cdot E_{\mu_t})^2 \tag{13}$$

By design, the orthogonal loss in (11) explicitly penalizes $R(F_{t-1}, E_{\mu_t})$, thereby reducing redundancy in the fused representation.

To interpret (12) more clearly, we rewrite it as:

$$\begin{aligned}I(F_t) = \quad & \underbrace{I(F_{t-1})}_{\text{current information}} \\ & + \underbrace{\left[I(E_{\mu_t} \mid F_{t-1}) - R(F_{t-1}, E_{\mu_t})\right]}_{\text{effective net new information}} \\ & + \underbrace{S(F_{t-1}, E_{\mu_t})}_{\text{synergy}}\end{aligned} \tag{14}$$

The term $I(E_{\mu_t} \mid F_{t-1}) - R(F_{t-1}, E_{\mu_t})$ represents the effective net new information contributed by E_{μ_t} beyond what is already captured in F_{t-1}. When $R(F_{t-1}, E_{\mu_t})$ decreases, this difference increases, thereby encouraging E_{μ_t} to contribute more complementary information.

Importantly, if F_{t-1} and E_{μ_t} together unlock new, synergistic insights, such synergy is not explicitly penalized by the loss function. Instead, the orthogonal loss selectively reduces redundancy while preserving (and potentially enhancing) meaningful joint interactions.

Finally, consider a sequence of N embeddings $\{E_{\mu_1}, E_{\mu_2}, \ldots, E_{\mu_N}\}$ fused step by step into $F_1, F_2, \ldots, F_N$. Reapplying (12) at each stage, the total information in F_N can be written as a telescoping sum:

$$\begin{aligned}I(F_N) = I(F_1) + \sum_{t=2}^{N} \big[I\big(E_{\mu_t} \mid F_{t-1}\big) - R\big(F_{t-1}, E_{\mu_t}\big) \\ + S\big(F_{t-1}, E_{\mu_t}\big)\big]\end{aligned} \tag{15}$$

where $F_1 = E_{\mu_1}$. The orthogonal loss (11) at each fusion step encourages $R(F_{t-1}, E_{\mu_t}) \to 0$. In that scenario, the cumulative sum (15) is dominated by the non-redundant and synergistic contributions:

$$\begin{aligned}I(F_N) \approx I(E_{\mu_1}) + \sum_{t=2}^{N} \big[I\big(E_{\mu_t} \mid F_{t-1}\big) \\ + S\big(F_{t-1}, E_{\mu_t}\big)\big]\end{aligned} \tag{16}$$

Hence, each modality contributes primarily complementary or synergistic information, rather than duplicating previous representations. Our experiments demonstrate the effectiveness of the OSF approach across various applications, as discussed in the following section.

IV. Experiments and Results

A. Datasets

We evaluate the versatility of OSF across various tasks using two multifaceted datasets.

1) CMU-MOSI: The CMU MOSI dataset [18] is a benchmark in multimodal sentiment analysis, featuring 2199 video segments combining three modalities: audio, visual, and textual data. It is widely employed for both fine-grained (7-class) and coarse-grained (2-class) sentiment classification, as well as regression tasks. The class labels and regression scores are derived from a linear scale that generates scores ranging from -3 to +3.

2) The Cancer Genome Atlas Kidney Renal Papillary Cell Carcinoma (TCGA-KIRP): The TCGA KIRP database [23] is commonly utilized for prediction tasks in kidney renal papillary cell carcinoma. From this database we build a dataset with up to six modalities, including images, clinical data, messenger ribonucleic acid (mRNA), micro ribonucleic acid (miRNA), deoxyribonucleic acid methylation (DNAm) and copy number variation. We use this dataset to demonstrate the applicability of our fusion approach in a highly multimodal, clinical setting.

B. Evaluation Metrics

1) 2-Class and 7-Class Classification: We use accuracy as the primary evaluation metric. Accuracy measures the model's capability to classify sentiment. However, it is important to note that the 7-class task inherently involves greater complexity, thus making high accuracy rates more significant.

2) Regression: For the regression task, we adopt Mean Absolute Error (MAE) as our metric. MAE measures the average absolute differences between the predicted and actual values, offering an intuitive understanding of the model's prediction quality in a continuous setting.

3) Survival Prediction: The concordance index (c-index) is used to evaluate the effectiveness of our model in survival analysis. It measures the ability of the model to correctly rank pairs of individuals based on their survival times. To further validate our model's reliability in survival prediction, we also calculate the Integrated Brier Score (IBS), which quantifies the prediction error over time.

C. Baselines

For the CMU MOSI dataset, we comply with data processing guidelines established by [24]. The primary objective is to evaluate the effectiveness of our OSF approach. We compare OSF against conventional multimodal fusion methods, including Early fusion, Late fusion, Mean fusion, Max fusion, and Sum fusion. Due to the simplicity of the models and the relatively low computational cost, we train separate models for each prediction task on the CMU MOSI dataset. Each modality, represented as a sequence, is initially processed by a Gated Recurrent Unit (GRU) with tanh activation. This is followed by a dense layer with rectified linear unit (ReLU) activation and we use a dropout rate of 0.1. This approach differs from the subsection IV-F, where a single model is trained based on MAE and all metrics are derived from this one model. For the TCGA-KIRP dataset, our data processing aligns with the approach described in [25]. The TCGA-KIRP experiments primarily use dense layers with ReLU activation for all modalities, except for the image data. The image modality is transformed into a feature vector using a pre-trained InceptionV3 [26] model with ImageNet weights [27]. Unlike the CMU MOSI experiments, a single model is trained for survival prediction, from which all metrics are derived.

Both datasets employ simple dense-layer models built on these encoded representations for each modality. The specific point at which fusion occurs depends on the baseline method in consideration.

D. Results and Implications

1) Results: Our experiments provide strong evidence of OSF's effectiveness across various tasks. Notably, OSF consistently outperforms traditional fusion techniques, delivering substantial improvements over existing methods. In the CMU-MOSI dataset, OSF emerges as the top-performing model across all three evaluation tasks, as detailed in Table I. For the 2-class classification, OSF achieves an accuracy of 0.687 ± 0.013, surpassing other standard methods. Similarly, OSF maintains its superior performance in both the 7-class classification and regression tasks. The lower standard deviations observed in performance metrics across all tasks indicate that OSF not only enhances accuracy but also ensures greater stability and reliability.

Our findings on the TCGA-KIRP dataset, detailed in Table II, further underscore OSF's strengths. Unlike other methods that hover around the randomness threshold in terms of c-index, OSF significantly exceed it. Specifically, the simple OSF model, built using the modality ordering from Section III-C, attains a c-index of 0.691 ± 0.100, clearly demonstrating its superior predictive capability. OSF also outperforms baseline methods in terms of IBS, further validating its robustness.

A key takeaway from Table II is the substantial performance gap between traditional fusion methods and OSF-based approaches. The simple OSF model already provides notable improvements over baseline fusion techniques, effectively integrating multimodal information. The advanced OSF model, developed using the ordering strategy from Section III-D, further enhances this process, achieving the highest c-index 0.703 ± 0.046 and the lowest IBS 0.198 ± 0.005. The results suggest that the advanced ordering and fusion strategy in OSF play a crucial role in improving predictive performance.

2) Implications and General Observations: The results are in perfect harmony with our initial hypotheses: an increase in the number of modalities and the inclusion of a custom loss function significantly enhance the performance of our model. This is particularly salient in the context of healthcare, where data is inherently multimodal, and traditional methods have often resorted to simplistic fusion mechanisms.

TABLE I
RESULTS ON THE CMU-MOSI DATASET (10 RANDOM SEEDS)

Model	2-classes	7-classes	Regression
Early fusion	0.627 ± 0.016	0.217 ± 0.042	1.298 ± 0.083
Late fusion	0.659 ± 0.017	0.282 ± 0.017	1.181 ± 0.030
Max fusion	0.661 ± 0.015	0.273 ± 0.014	1.270 ± 0.034
Mean fusion	0.659 ± 0.018	0.284 ± 0.011	1.214 ± 0.024
Sum fusion	0.645 ± 0.017	0.278 ± 0.019	1.304 ± 0.114
OSF	$\mathbf{0.687 \pm 0.013}$	$\mathbf{0.300 \pm 0.015}$	$\mathbf{1.139 \pm 0.023}$

E. Exhaustive Permutations Analysis

We validated the proposed heuristic for modality ordering by exhaustively evaluating all $N!$ permutations of the modalities. For $N = 6$, this required training and evaluating $6! = 720$ models. While such an exhaustive approach is computationally

TABLE II
RESULTS ON THE TCGA-KIRP DATASET (10 RANDOM SEEDS)

Model	c-index	IBS
Early fusion	0.529 ± 0.006	0.484 ± 0.159
Late fusion	0.528 ± 0.005	0.475 ± 0.147
Max fusion	0.531 ± 0.001	0.500 ± 0.132
Mean fusion	0.531 ± 0.001	0.374 ± 0.159
Sum fusion	0.530 ± 0.001	0.406 ± 0.135
OSF (simple)	$\mathbf{0.691 \pm 0.100}$	$\mathbf{0.213 \pm 0.005}$
OSF (advanced)	$\mathbf{0.703 \pm 0.046}$	$\mathbf{0.198 \pm 0.005}$

infeasible for larger N, this analysis provides critical insights into the performance landscape and establishes a benchmark for assessing the heuristic's efficacy.

For each permutation, a model was trained and evaluated using the c-index (c) and IBS (b) as performance metrics. Since these metrics follow opposite optimization directions—where higher values of c indicate better performance, while lower values of b are preferred—we adjust Δ_i accordingly:

$$\Delta_i = (\tilde{c}_i - \sigma_i^c) - (\tilde{b}_i + \sigma_i^b) \tag{17}$$

Similarly, for bimodal models, the corresponding pairwise performance metric is:

$$\Delta_{ij} = (\tilde{c}_{ij} - \sigma_{ij}^c) - (\tilde{b}_{ij} + \sigma_{ij}^b) \tag{18}$$

This formulation ensures that Δ favors configurations with high discrimination (c-index) and low prediction error (IBS), allowing the heuristic to identify robust unimodal and bimodal orderings efficiently.

The performance landscape, shown in Fig. 2, reveals a gradient in the c-index vs. IBS space, emphasizing the critical role of selecting an effective modality ordering. Models with both low IBS and low c-index are absent, consistent with theoretical findings in survival analysis [28]. The IBS accounts for both calibration and discrimination, whereas the c-index evaluates discrimination alone. Models near the Pareto-optimal boundary achieve balanced performance, combining high c-index and low IBS.

As shown in Fig. 2, the heuristic-derived ordering lies among the top-performing models, demonstrating its ability to approximate optimal orderings without requiring exhaustive computations. Its placement near the Pareto-optimal boundary in the c-index vs. IBS space (Fig. 2) highlights its computational efficiency and experimental foundation, making it a reliable tool for modality ordering in OSF.

F. Model Integration

To demonstrate the adaptability and effectiveness of our proposed method, we integrated it into the Multimodal Mutual Information Maximization (MMIM) framework [29], a well-established model noted for its strong performance on the CMU-MOSI dataset. Our objective was twofold: to ascertain the ease of integrating our method into a fully developed and optimized model, and to evaluate its performance impact.

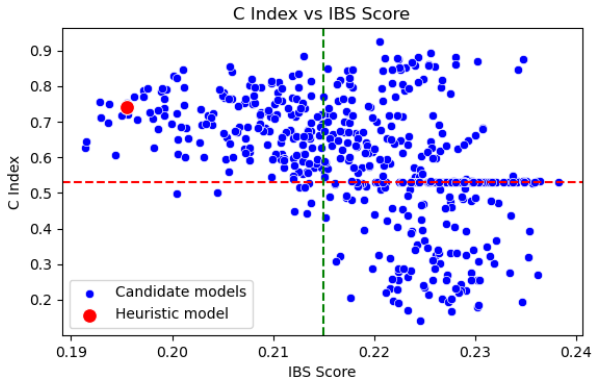

Fig. 2. c-index vs. IBS performance landscape for all 720 permutations. The top-left corner represents Pareto-optimal models. The heuristic-derived model is highlighted and lies among the top-performing models, validating the efficacy of the proposed heuristic.

In replicating the MMIM framework, we adhered to the same data processing methods, experimental setup, and other implementation details as in the original paper. OSF integration into MMIM specifically occurred at the level of the Fusion Network. In our approach, we modified this step to sequentially merge modalities without any major changes to the MMIM architecture. For OSF, the contrastive part of the conventional MMIM was removed not to interfere with our orthogonal loss. This led to two notable impacts: first, an improvement in performance; and second, a reduction in training time per epoch from 75s to 45s on a NVIDIA©T4 GPU.

TABLE III
RESULTS ON THE CMU-MOSI USING MMIM FRAMEWORK (20 RANDOM SEEDS)

Model	2-classes	7-classes	Regression
MMIM	0.826 ± 0.006	0.458 ± 0.012	0.723 ± 0.013
MMIM + OSF	$\mathbf{0.831 \pm 0.005}$	$\mathbf{0.462 \pm 0.016}$	$\mathbf{0.719 \pm 0.015}$

TABLE IV
RESULTS ON THE CMU-MOSI USING MMIM FRAMEWORK (BEST RESULTS)

Model	2-classes	7-classes	Regression
MMIM	0.838	0.477	0.702
MMIM + OSF	**0.843**	**0.488**	**0.682**

Table III and IV presents the results of the MMIM framework and its variations with OSF. OSF exhibited superior performance across all evaluation metrics, both in terms of mean and best values.

These findings underscore not only the ease with which our proposed methods can be integrated into existing frameworks but also their potential to improve overall model performance—especially in computational efficiency, as evidenced by the reduced training time for OSF but also in terms of evaluation metrics.

V. Conclusion

In this paper, we introduce Orthogonal Sequential Fusion (OSF), a new fusion paradigm for multimodal machine learning. Our approach relies on two major contributions: a sequential mechanism for processing modalities and an orthogonal loss function. The sequential mechanism provides an unprecedented level of control over modality combination, enabling a precise representation of intrinsic intermodal relationships. The orthogonal loss function, through its exploitation of data complementarities and redundancies, amplifies the extraction of valuable insights from the multimodal context. Collectively, these innovative components establish OSF as a notable alternative to traditional fusion techniques in multimodal machine learning.

Throughout comprehensive experimental testing, OSF consistently outperforms traditional fusion strategies such as early fusion, late fusion, and mean/max/sum fusion. It differentiates itself not only with superior performance, but also with a more adaptable approach. This blend of flexibility and performance solidifies OSF as a viable and promising alternative to existing fusion paradigms. The empirical evidence underscores OSF's potential to significantly enhance the performance of multimodal machine learning models, demonstrating a unique combination of adaptability and improved model performance.

While our approach's promising results underscore its effectiveness, we acknowledge certain limitations and potential avenues for future work. Key challenges include the scalability of our method to a larger number of modalities, its manageability given the high number of hyperparameters, and the prospect of exploring alternative orthogonalization techniques.

VI. Rights and Permissions

For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

References

[1] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer vision and image understanding*, vol. 108, no. 1-2, pp. 116–134, 2007.

[2] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, "Multimodal end-to-end autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 537–547, 2020.

[3] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[4] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 944–10 956, 2021.

[5] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[6] G. Barnum, S. Talukder, and Y. Yue, "On the benefits of early fusion in multimodal representation learning," *arXiv preprint arXiv:2011.07191*, 2020.

[7] G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*. Springer, 2011, pp. 396–406.

[8] Z.-z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann, "Multimedia classification and event detection using double fusion," *Multimedia tools and applications*, vol. 71, pp. 333–347, 2014.

[9] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[10] 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*, vol. 30, no. 12, pp. 4467–4480, 2019.

[11] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *Advances in neural information processing systems*, vol. 33, pp. 4835–4845, 2020.

[12] A. M. Saxe, J. L. McClelland, and S. Ganguli, "A mathematical theory of semantic development in deep neural networks," *Proceedings of the National Academy of Sciences*, vol. 116, no. 23, pp. 11 537–11 546, 2019.

[13] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, "Reducing overfitting in deep networks by decorrelating representations," *arXiv preprint arXiv:1511.06068*, 2015.

[14] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[15] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[16] 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*. Springer, 2021, pp. 667–677.

[17] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty." *science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[18] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[19] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[20] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying unique information," *Entropy*, vol. 16, no. 4, pp. 2161–2183, 2014.

[21] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions," *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–42, 2024.

[22] N. Bansal, X. Chen, and Z. Wang, "Can we gain more from orthogonality regularizations in training deep networks?" *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[23] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (tcga): an immeasurable source of knowledge," *Contemporary oncology*, vol. 19, no. 1A, p. A68, 2015.

[24] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu *et al.*, "Multibench: Multiscale benchmarks for multimodal representation learning," *Advances in neural information processing systems*, vol. 2021, no. DB1, p. 1, 2021.

[25] L. A. Vale-Silva and K. Rohr, "Long-term cancer survival prediction using multimodal deep learning," *Scientific Reports*, vol. 11, no. 1, p. 13505, 2021.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[28] S. Y. Park, J. E. Park, H. Kim, and S. H. Park, "Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches)," *Korean Journal of Radiology*, vol. 22, no. 10, p. 1697, 2021.

[29] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," *arXiv preprint arXiv:2109.00412*, 2021.