International Journal of Internet, Broadcasting and Communication Vol.17 No.3 144-154 (2025)
http://dx.doi.org/10.7236/IJIBC.2025.17.3.144

IJIBC 25-3-17

MRO: Multimodal Routing Optimization via Neural Architecture Search of Fusion Paths

Jeong-Hun Kim, Mi-Hwa Song*

Undergraduate Researcher, Division of Computer Engineering, Hansung University
*Associate Professor, Division of Computer Engeneering, Hansung University**
*bandlagom0927@hansung.ac.kr, mhsong@hansung.ac.kr**

Abstract

Effective fusion of heterogeneous modalities is a critical factor for improving performance in multimodal learning. However, existing Cross-Modal Transformer (CMT)-based fusion methods are constrained by fixed attention paths, which reduce adaptability to diverse inputs and restrict flexible exploration of optimal fusion strategies. Furthermore, as the number of modalities increases, the computational complexity grows exponentially, leading to scalability bottlenecks. To address these limitations, we propose a Multimodal Routing Optimization (MRO) framework that restructures cross-modal attention paths as a Supernet structure, drawing inspiration from the Once-for-All (OFA) paradigm. Leveraging Neural Architecture Search (NAS), Multimodal Routing Optimization (MRO) dynamically selects optimal routing paths that balance accuracy and computational cost (FLOPs), enabling scalable and efficient multimodal fusion.

Keywords: *Neural Architecture Search, Multimodal Learning, Fusion Paths, Multimodal Routing Optimization (MRO)*

1. Introduction

The fusion of several different types of information is central to multimodal learning. These types of information include text, voice and images. This field strives to achieve a more profound comprehension. It does this by integrating these diverse data sources. The Cross-Modal Transformer (CMT)[1] architecture is used across many research projects to deal with complex interactions between different types of data. CMT allows for precise modeling of cross-modal semantic transitions by assigning one modality as the query and performing attention over the others.

Two main approaches to the existing CMT-based fusion method are identified. The first is a Pairwise Full Attention structure that performs cross-modal attention across all pairs of input modalities. The former offers high expressive power, but as the number of modalities (k) increases, the computational cost scales quadratically ($O(k^2)$), leading to bottlenecks. Conversely, while the latter can reduce computation time,

Manuscript Received: July. 2, 2025 / Revised: July. 15, 2025 / Accepted: July. 15, 2025
Corresponding Author: mhsong@hansung.ac.kr
Tel: +82-2-760-4124, Fax: +82-2-760-5771
Associate Professor, Division of Computer Engineering, Hansung University, Korea*

its fixed path design means it is unable to respond flexibly to changes in data quality or input diversity, and it is prone to performance instability.

To solve this problem, Multimodal Routing Optimization (MRO) treats attention paths as search variables, expands the Pairwise CMT with all cross-modal paths into a Supernet, and defines it as a searchable structure using binary masks applied to each attention path. It then selects the optimal subnet by considering the trade-off between accuracy and computational cost (FLOPs) through Neural Architecture Search (NAS). This study combines existing technologies, such as Supernet design, Prior-Guided NAS, and Graph Attention.

However, it is original in that it proposes a new theoretical framework that redefines the "multimodal attention path selection problem" as a discrete optimization problem rather than simply aggregating these technologies. Unlike previous NAS-based studies, which explored the overall model structure or fusion strategy, this study treats each cross-modal attention path as a binary variable. The study then constructs an exploration space at the path level by combining paths to select the optimal subnet. This structure enables the quantitative adjustment of the balance between semantic validity, computational efficiency, and performance. This is achieved by assigning prior probabilities through a meaning-based modal relation graph and enforcing consistency between prior expectations and actual performance via rank regularization. This forms an interpretable, structured NAS framework. Thus, the theoretical contribution of this study lies in defining a new search space and corresponding optimization structure for multimodal path selection that goes beyond the combination of existing techniques.

2. Related Works

2.1 Neural Architecture Search (NAS)

The Neural Architecture Search (NAS)[2] method is comprised of the following three factors. First, neural networks are explored automatically. Second, there is a navigation algorithm. Third, there is a performance evaluation algorithm. In the early stages, network architectures are sampled repeatedly through a reinforcement learning–based controller. Subsequently, NASNet was successfully applied to image classification. Since the publication of that seminal paper, various evolutionary algorithms have been developed, and significant advances have been made in their practical application. This study aims to convert the path selection process to explore it, as opposed to using a fixed multidimensional fusion or multi-modal convergence structure.

2.2 Lightweight Supernet-based NAS: Once-for-All (OFA)

The Once-for-All (OFA)[3] concept proposes a flexible framework to support different neural network configurations within one large-scale supernet. This approach has been proven to be efficient in that it utilizes masking techniques to enable rapid evaluation and deployment of substructure networks through masked one-shot learning without the need for retraining. Building on the core principles of OFA, this study extends the Pairwise CMT model into a supernet structure to explore effective routing path combinations suitable for various multimodal contexts. Such a design can adaptively support multimodal convergence even in situations with limited data availability, and at the same time, it is a promising way to improve computational efficiency

2.3 Prior-guided Constrained Search

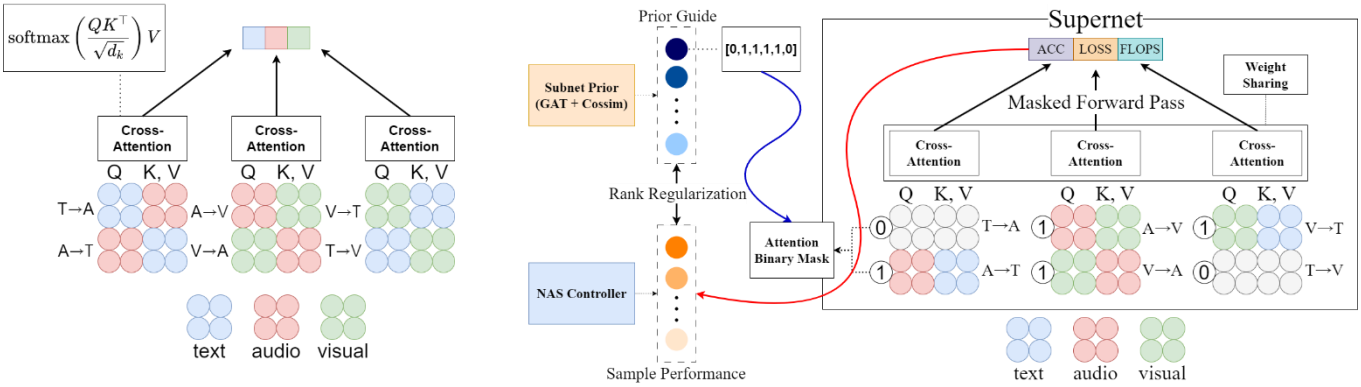
The issue of expanding the navigation space has been the focus of recent research in the field of NAS. This is to facilitate effective navigation. It is also to establish appropriate search efficiency and structural properties. One notable study involves PGONAS [4]. The alignment between the subnet candidate group ranking and the subsequent performance of the subnet candidate group and the actual performance is demonstrated by this.

Meanwhile, GraphCFC [5] formalizes modality interactions as conditionally connected graphs, using them as prior

distributions to guide early NAS exploration. This approach prioritizes meaningful subnetwork configurations, improving search efficiency by avoiding random, uninformative sampling.

3. MRO : Multimodal Routing Optimization

3.1 MRO Overview



(A) Pairwise Cross Modal Transformer (B) MRO Cross Modal Transformer

Figure 1. MRO Structure

As illustrated in Figure 1, the architectural differences between the baseline Pairwise Cross-Modal Transformer (CMT) and the proposed Multimodal Routing Optimization (MRO) framework are evident.

(A) illustrates the Pairwise CMT architecture, the pairwise attention scheme proposed in One-vs-Other [6]. The model performs all the cross-attention paths (T→A, T→V, A→T, A→V, V→T, and V→A) between all modality pairs, exhibiting high levels of expressive power. However, as the number of modalities increases, the amount of computation increases rapidly and a problem of lack of flexibility due to restructuring occurs.

(B) illustrates the proposed MRO framework, which expands the Pairwise CMT into a Supernet and selectively controls whether each path is activated through the Attention Binary Mask. MRO uses a GAT (Graph Attention Network)[7]-based directional graph and cosine similarity to measure the importance of interactions between modalities and create a Prior Guide. The purpose of the Prior Guide is to select more promising subnet configurations during the initial exploration phase of the NAS. The selected subnets are attention masked to fix their structure, and the NAS controller updates its parameters based on the performance of those masked subnets. Both the evaluation and granularity are controlled during this process.

In this process, rank regularization between prior and sample performance is applied. This is an important step in minimizing prior-based search bias and selecting subnets that match actual performance more closely. The proposed approach is designed to optimize the computational efficiency of the learning process, utilizing a masked one-shot learning method with weight sharing within Supernet.

3.2 Supernet Architecture for Cross-Modal Attention Path Exploration

The proposed framework integrates all pairwise attention routes among three modalities – text, audio, and vision – into a unified Cross-Modal Transformer (CMT) Supernet. Specifically, six cross-modal attention paths (T→A, T→V, A→T, A→V, V→T, V→A) are each implemented as independent attention blocks within the Supernet, and their activation is controlled via a binary mask vector.

Each constituent element of the Supernet is made up of standard Transformer components, which have been implemented in PyTorch. These components include multi-head self-attention, layer normalization and a

feedforward network. Linear transformation layers are used to project each input modality into a fixed embedding space. During the forward pass, solely the attention blocks corresponding to active mask entries (i.e. mask value = 1) are activated and executed, while inactive paths are wholly bypassed to reduce computational overhead. Consider, for example, a mask vector of [1, 0, 1, 1, 0, 1]. This activates four out of six possible attention paths, and it is only those that contribute to the fusion process. The outputs from the active paths are aggregated via average pooling and passed to a shared classification head for final prediction. A weight-sharing paradigm is used to train the Supernet to facilitate efficient neural architecture search (NAS). In this paradigm, all possible subnets reuse the same parameters, meaning retraining is not necessary. This facilitates rapid evaluation of multiple subnet configurations during the NAS process. It does this by significantly reducing computational costs. At the same time, it maintains structural flexibility.

Table 1. Pseudocode for NAS-Based Exploration of Cross-Modal Transformer Paths

```
Supernet ← build_CMT_Supernet()
Prior ← compute_initial_prior()

for epoch in range(E):
    if epoch < Warmup:
        Masks ← uniform_sample(n)
    else:
        Masks ← prior_guided_sample(Prior, n)

    Losses ← []
    for mask in Masks:
        Subnet ← Supernet(mask)
        L ← TaskLoss(Subnet)
        optimize(L)
        Losses.append(L)

    R ← 0
    for (i, j) in pairwise(Masks):
        R += max(0, (Losses[i] - Losses[j]) * sign(Prior[Masks[i]] - Prior[Masks[j]]))
    optimize(R)

    Prior ← update_prior(Masks, Losses)

TopK ← select_final_subnets(Prior, Logs)
```

The neural architecture search (NAS) process for identifying optimal attention path configurations is illustrated in Table 1. Initially, a prior distribution over all attention paths is computed based on semantic similarity scores derived from the Modal Relation Graph. During the warm-up phase, Attention Binary Masks are sampled uniformly to ensure diverse exploration of the search space. Once warm-up ends, the algorithm transitions to a prior-guided sampling strategy, where subnet candidates are sampled with probabilities biased by the learned prior. Each sampled mask activates a specific subset of attention paths within the Supernet, defining a distinct Subnet architecture. For each Subnet, task-specific loss (e.g., cross-entropy for classification) is computed and backpropagated, updating the shared weights in the Supernet accordingly. To maintain alignment between the prior expectation and observed Subnet performance, a rank regularization term is introduced to encourage consistency in relative ranking. This term penalizes mismatches between the predicted ranking (from the prior) and the observed ranking (based on loss values) among sampled Subnets. The regularization loss is added to the overall objective and optimized jointly.

After each epoch, the prior distribution is updated using the observed performance of the sampled Subnets. This iterative update process helps guide the search toward promising regions of the architecture space. Once the search concludes, the top-k Subnets with the highest prior scores and/or best validation performance are selected for final evaluation or deployment. This masked one-shot NAS strategy allows the proposed MRO framework to dynamically identify efficient and high-performing attention configurations without retraining or exhaustive search, making it scalable and practical for real-world multimodal applications.

3.3 Prior-Guided Search Using a Modal Relation Graph

The proposed study introduces a Modal Relation Graph (MRG) to reduce the search space semantically and prioritize critical modality interactions during multimodal architecture exploration. The MRG is defined as a directed graph where each node represents a modality, and each edge encodes the strength of interaction between modalities. The importance of each edge is computed by aggregating three statistical metrics: semantic similarity, gradient sensitivity, and representational diversity.

$$p_{i\to j} = \frac{\exp\Big(\mathrm{GAT}\Big(\alpha\cdot\mathrm{CosSim}(i,j)+\beta\cdot\mathrm{Var_g}(i,j)+\gamma\cdot\mathrm{Var_a}(i,j)\Big)\Big)}{\sum_{k\neq i}\exp\Big(\mathrm{GAT}\Big(\alpha\cdot\mathrm{CosSim}(i,k)+\beta\cdot\mathrm{Var_g}(i,k)+\gamma\cdot\mathrm{Var_a}(i,k)\Big)\Big)} \tag{1}$$

Specifically, for a given modality pair i and j, three indicators are computed to quantify the edge weight from modality i to modality j in the Modal Relation Graph (MRG). First, the semantic similarity between modalities i and j is measured by computing the cosine similarity of their embedding vectors. This reflects the degree of alignment between the two modalities in the representational space. Second, the gradient variance between modalities i and j indicates the extent to which the corresponding attention path contributes to model parameter updates during training. A higher variance suggests a more dynamic and influential role in optimization, highlighting the path's relative importance. Third, the attention activation variance between modalities i and j quantifies the fluctuation in attention weights along the path, capturing both the diversity and stability of inter-modal interactions. This provides insight into the informativeness and robustness of the modality pair.

These three metrics are linearly combined with trainable weights alpha, beta, and gamma to produce a composite score for each directed edge. This composite score is then modulated via a Graph Attention Network (GAT) to incorporate the contextual relevance of each modality within the graph structure. Finally, the attention scores from modality i to all other modalities j (where $j \neq i$) are normalized via a softmax function, yielding a prior probability distribution over attention paths. This prior distribution guides the NAS controller during architecture search. Consequently, attention paths with higher semantic and functional relevance are more frequently sampled, improving search efficiency and enhancing the model's overall representational capacity.

3.4 Prior-Guided Search with Rank Regularization for Consistency Alignment

To enhance the efficiency and precision of the search process, the proposed framework employs a prior-guided NAS strategy. This approach leverages semantically grounded prior knowledge to reduce the reliance on random architecture sampling. Specifically, the importance of each attention path is estimated using the Modal Relation Graph (MRG), which encodes semantic similarity and statistical interactions among modalities. These scores are transformed into a probability distribution via softmax normalization and used as a prior to guide the selection of subnet configurations. However, since the prior does not always perfectly correlate with actual performance, inconsistencies can arise during the search process. To mitigate this issue, we introduce a rank regularization loss that aligns the sampling prior with empirical performance. This auxiliary loss is designed to minimize the discrepancy between the ranking order induced by the prior and the actual task-specific performance of sampled subnets in each NAS round. The rank regularization loss is formally defined as:

$$\mathcal{L}_{\text{rn}} = \frac{1}{N} \sum_{i=1}^{N} \left| \text{rank}_i^{\text{prior}} - \text{rank}_i^{\text{perf}} \right|^p \tag{2}$$

The value of p determines the type of norm used, and we empirically explored p in $\{1, 2, \infty\}$ in our ablation study. We also evaluated an additional formulation based on the L0 norm, which simply counts the number of mismatches in ranking positions between the prior and performance orders. Among all variants, the L2-based version generally provided the best balance between accuracy and ordinal consistency. The optimal choice of p was determined individually for each dataset based on a trade-off between Spearman rank correlation and predictive accuracy. we selected the most suitable norm type for each case accordingly (see Section 4.3 for details). where rank_prior_i and rank_perf_i denote the prior-based and performance-based rankings of the i-th sampled subnet, respectively, and N is the number of subnet samples evaluated in the current search iteration. Spearman rank correlation is also computed and monitored to preserve ordinal consistency. This dynamic alignment improves the reliability of the prior distribution over time, guiding the NAS controller toward subnet architectures that are semantically meaningful and empirically effective. The architecture space becomes increasingly refined, ultimately yielding subnet configurations that strike a favorable balance between high predictive accuracy and computational efficiency.

4. Evaluation

4.1 Learning Dataset

Table 2. Distribution of 7-Class Emotion Labels in CMU-MOSEI and MELD

Dataset	CMU_MOSEI	MELD
sentiment	Mean Probability	Mean Probability
Happy (Joy)	27.76%	15.8%
Sadness	32.23%	9.0%
Angry	14.61%	14.3%
Disgust	10.51%	3.0%
Fear	3.4%	2.5%
Surprise	6.49%	11.3%
Neutral	5.0%	44.1%
total	100%	100%

This study independently used CMU-MOSEI [8] and MELD [9], public multimodal emotion recognition datasets, in the experiment to verify the performance and lightweight effect of the proposed Multimodal Routing Optimization (MRO) structure from various angles. The two datasets complement each other in terms of their composition, speech units, modal combinations and labelling systems, making them suitable for evaluating the performance of MRO under various input conditions. CMU-MOSEI consists of approximately 23,500 single utterances from individual speakers, with each sample containing text, voice, and facial expression, and is labelled with seven emotion classes. The structure is based on a single speaker, and the time consistency and interaction between modalities are clear. This provides a basic structure verification environment for MRO to maintain performance while reducing the amount of computation. MELD is a multimodal dataset consisting of over 13,000 utterances from the TV show Friends, aligned across text, speech, and visual modalities. It records conversations between speakers, how they flow, and how emotions change. This makes it a good way to measure choosing paths and understanding the situation, which are important parts of MRO. Distinct modality characteristics are possessed by CMU-MOSEI and MELD, so Attention Binary Mask optimization is performed separately, and the two datasets are trained independently without parameter sharing. This allows for a systematic validation of MRO's robustness and adaptability.

4.2 Model Hyperparameter

Table 3. Hyperparameter Settings for MRO on the CMU-MOSEI and MELD Datasets

Dataset	CMU_MOSEI	MELD
Batch size	32	64
epoch	40	30
Learning rate	3e-5	5e-5
Dropout	0.4	0.3
Model dim	256	256

The model hyperparameters were configured by considering each dataset's sample size, utterance length, and emotional label distribution to ensure both training stability and convergence speed. Although CMU-MOSEI and MELD share similar modality compositions, they differ significantly in terms of scale and contextual complexity, necessitating distinct hyperparameter settings even under the same architecture.

For CMU-MOSEI, which consists of over 23,000 samples with relatively long utterances and strong temporal alignment across modalities, a batch size of 32 was chosen. The model was trained for 40 epochs to allow sufficient learning over the extensive data. A learning rate of 3e-5 was adopted to promote stable convergence, and a dropout rate of 0.4 was applied to prevent overfitting. The multimodal embedding dimension was fixed at 256 to maintain computational stability and consistency within the CMT-based architecture. In contrast, MELD is a conversational multimodal dataset characterized by frequent context shifts, shorter utterances, and a smaller sample size of approximately 13,000. To improve training efficiency under these conditions, a larger batch size of 64 was used, and the number of training epochs was reduced to 30 to mitigate overfitting. The learning rate was set to 5e-5 to encourage faster convergence, while a dropout rate of 0.3 was applied to support generalization performance.

4.3 Ablation Study on Rank Regularization Loss Functions

To evaluate the effectiveness of the rank regularization strategy in aligning the consistency between prior-based and performance-based rankings, ablation experiments were conducted using different loss formulations. Specifically, the rank discrepancy was computed under three variants: L1 (absolute error), L2 (squared error), and L∞ (maximum error). All experiments were carried out under identical NAS conditions, with only the ranking loss function varied while keeping the remaining architectural components fixed.

Table 4. Ablation Study: Impact of Rank Regularization Loss Functions

Dataset	Loss Function	Attention Mask	Acc-7	Acc-2	F1	Rank Corr (ρ)
CMU_MOSEI	L0	[1,0,0,0,1,0]	51.23	83.79	78.64	0.356
	L1	[0,1,1,0,1,1]	55.30	86.45	86.30	0.613
	L∞	[1,1,1,0,1,1]	52.66	85.50	81.11	0.237
	L2	[0,1,1,0,1,1]	55.30	86.45	86.30	0.482
MELD	L0	[1,0,0,1,1,0]	64.52	88.42	62.30	0.462
	L1	[0,1,1,1,1,1]	66.45	89.67	65.40	0.285
	L∞	[1,1,0,0,1,1]	65.30	88.90	63.20	0.071
	L2	[0,1,1,1,1,0]	67.07	90.32	66.88	0.318

Table 4 compares the performance of subnet selection when using different rank regularization loss functions (L0, L1, L2, and L∞) on the CMU-MOSEI and MELD multimodal emotion recognition datasets. The evaluation metrics that will be used are Acc-7, Acc-2, F1-score, and the Spearman rank correlation coefficient (ρ). The latter measures

the ordinal consistency between prior-based and performance-based rankings.

The CMU-MOSEI dataset has a relatively balanced distribution of emotions, requiring a model to account for a wide range of attention path combinations to maximize performance. In this setting, L1 loss achieved the highest rank correlation ($\rho = 0.613$) and produced strong results across all metrics. Acc-7 = 55.30%, Acc-2 = 86.45%, and F1-score = 86.30%. This suggests that L1 effectively moderates rank differences without imposing excessive penalties, enabling the precise selection of semantically meaningful paths. Conversely, MELD exhibits high emotional skew, with the neutral class comprising nearly 45% of the data. This imbalance increases the sensitivity of model performance to specific paths. Under these conditions, stable path selection is more important than rank alignment, and L2 loss is most suitable. Although its rank consistency ($\rho = 0.318$) was relatively low, it achieved the best overall performance. Acc-7 = 67.07%, Acc-2 = 90.32%, and F1-score = 66.88%.

In conclusion, L1 loss was the most suitable choice for CMU_MOSEI because it encouraged exploring meaningful attention paths while maintaining high rank alignment. For MELD, however, L2 loss was more effective because it optimized performance directly under conditions of strong class imbalance. Based on the characteristics of these datasets, all subsequent NAS and fine-tuning experiments in this study adopted L1 loss for CMU_MOSEI and L2 loss for MELD.

4.4 Performance Results

We evaluate the performance of four representative multimodal fusion architectures in our experiments. First, the Concat + MLP approach simply concatenates the outputs from each modality and feeds them into a multilayer perceptron for classification. Second, EmbraceNet is designed to perform robust predictions even when some modality information is missing. Third, the Cross-Modal Transformer (CMT) learns all pairwise attentions across modality combinations, offering high representational power but incurring significant computational cost. Lastly, the proposed Multimodal Routing Optimization (MRO) builds upon the full CMT Supernet by including all attention paths and applies a prior-guided NAS strategy to optimize the trade-off between accuracy and computational efficiency.

Table 5. Performance and FLOPs of MRO and Baselines on CMU-MOSEI and MELD

Dataset	Model	Subnet Mask	MAE	Corr	Acc-7	Acc-2	F1	MFLOPs
CMU_MOSEI	Concat+MLP	-	0.645	0.601	41.20	77.85	77.40	22.31
	EmbraceNet	-	0.602	0.644	44.30	79.10	78.95	26.78
	Pairwise CMT	[1,1,1,1,1,1]	0.558	0.721	50.90	84.20	83.80	254.94
	MRO w/o Prior Guidance	[1,1,0,0,1,0]	0.591	0.706	49.16	82.74	80.92	142.53
	MRO w/o MRG Prior	[1,1,1,0,1,1]	0.546	0.735	52.66	85.50	81.11	190.72
	MRO w/o Rank Regularization	[1,1,0,0,1,1]	0.530	0.758	53.69	85.91	85.0	178.41
	Full MRO	[0,1,1,0,1,1]	0.519	0.763	55.30	86.45	86.30	173.48
MELD	Concat+MLP	-	0.634	0.589	61.83	85.67	61.21	23.90
	EmbraceNet	-	0.598	0.631	63.92	87.04	63.57	29.20
	Pairwise CMT	[1,1,1,1,1,1]	0.561	0.683	66.14	89.36	65.76	378.82
	MRO w/o Prior Guidance	[1,1,1,1,1,0]	0.614	0.673	61.20	86.73	63.50	331.53
	MRO w/o MRG Prior	[0,1,1,0,1,1]	0.592	0.682	64.92	88.04	64.87	299.62
	MRO w/o Rank Regularization	[0,1,1,1,1,1]	0.565	0.687	66.45	89.67	65.40	315.79
	Full MRO	[0,1,1,1,1,0]	0.559	0.691	67.07	90.32	66.88	297.36

Table 5 presents a comparative analysis of model performance metrics—MAE, Corr, Acc-7, Acc-2, and F1—and computational cost in MFLOPs on the CMU-MOSEI and MELD datasets. The results show that the proposed MRO consistently outperforms Pairwise CMT in accuracy metrics (Acc-7, Acc-2, F1) on both datasets, while significantly

reducing computational cost by selectively activating a subset of attention paths. For instance, on CMU-MOSEI, MRO achieves an Acc-7 of 55.30%, 4.4 percentage points higher than CMT, with a 32% reduction in FLOPs. A similar trend appears on MELD, with a 0.93% Acc-7 gain and 21% lower FLOPs.

To clarify the contribution of each component in MRO, three ablation variants are included:
(1) *MRO w/o Prior Guidance* removes guided sampling and uses random exploration,
(2) *MRO w/o MRG Prior* disables the semantic modal graph while retaining rank alignment,
(3) *MRO w/o Rank Regularization* omits the alignment between prior and performance rankings.
Each component proves beneficial, with full MRO achieving the best balance. Removing any component leads to performance drops—especially in Corr and Acc-2—highlighting the importance of their synergy.

The accuracy differences across datasets stem from label structure. CMU-MOSEI uses soft labels (e.g., Neutral: 0.6, Sad: 0.3), leading to ambiguous samples and lower accuracy but higher F1 due to balanced predictions. MELD, with discrete labels, shows higher Acc-7/Acc-2 but lower F1 due to class imbalance.

In summary, MRO uses a GAT-based semantic prior and cosine similarity to guide selective attention path activation. Through NAS optimization, it builds efficient subnetworks that preserve expressiveness, achieving robust multimodal fusion under both soft and hard label settings.

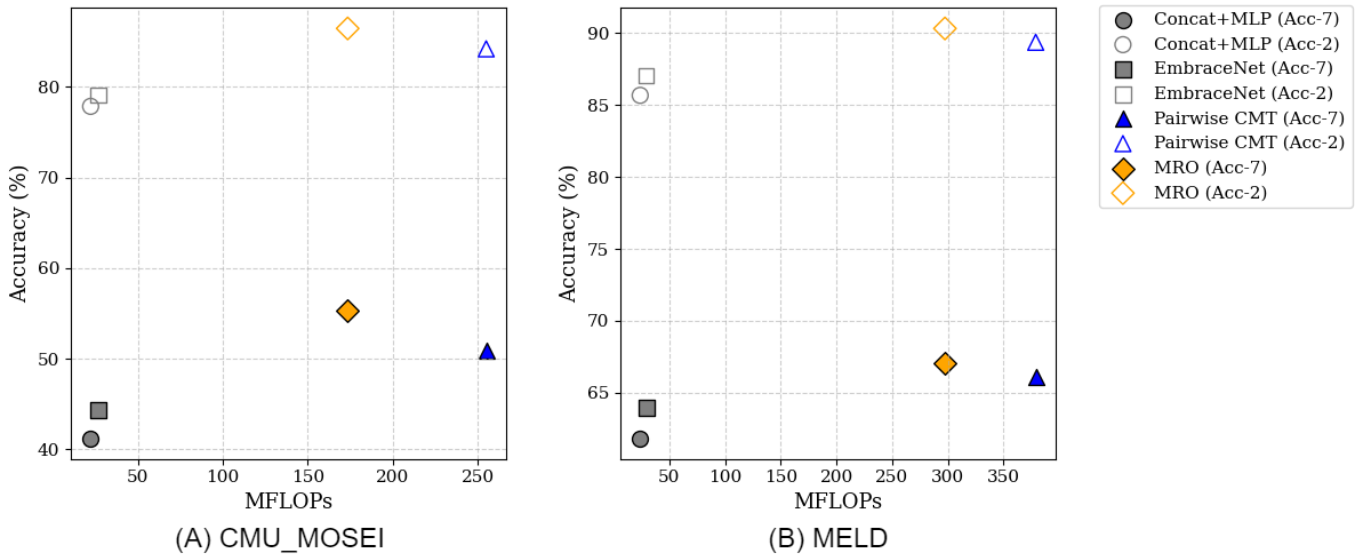


Figure 2. Accuracy – FLOPs Tradeoff

Figure 2 shows a trade-off curve visualizing the accuracy of each multimodal fusion model (Acc-7 and Acc-2) according to the amount of computation (MFLOPs) based on the CMU-MOSEI (left) and MELD (right) datasets. The CMU-MOSEI-based experimental results presented on the left show that basic structures such as Concat+MLP and EmbraceNet perform relatively poorly in terms of prediction accuracy, despite the advantage of low computation. The experimental results show that the Pairwise CMT model, which comprehensively utilizes cross-modal attention paths between all modalities, achieved high accuracy but has the drawback of very high computational complexity.

MRO achieves better performance and computational efficiency than traditional CMT models. It achieves slightly higher accuracy than CMT on both the Acc-7 and Acc-2 criteria with fewer concurrent computations (MFLOPs). Specifically, on the MELD dataset, MRO saves approximately 21% of FLOPs, even though the emotion classification accuracy (Acc-2) is almost identical. These results demonstrate that, rather than using all paths, MRO efficiently utilizes computational resources by selectively activating only meaningful attention paths. It remains expressive while eliminating unnecessary paths, thus increasing efficiency without compromising performance.

As can be seen in Figure 2, MRO is located at the top left of the accuracy-computation balance, providing visual confirmation of Pareto efficiency. Beyond numerical comparisons, this helps us to intuitively understand the trade-off between performance and efficiency in different structures.

5. Conclusions

To address the limitations of existing multimodal fusion models—namely, their reliance on fixed cross-modal attention paths that lack adaptability to diverse input conditions and resource constraints—we propose Multimodal Routing Optimization (MRO). MRO structures the Pairwise Cross-Modal Transformer (CMT) as a Supernet encompassing all possible modality pairwise attention paths and uses an attention binary mask to dynamically activate different subnet configurations. To enhance search efficiency, we introduce a Prior Guide based on semantic similarity and interaction strength among modalities. Building on the Once-for-All paradigm, we implement a masked one-shot learning scheme with weight-sharing, ensuring both computational efficiency and training stability. A key innovation is the Rank Regularization strategy, which aligns the ranking of prior probabilities with actual subnet performance, effectively reducing bias in NAS-driven architecture selection. Extensive experiments on the MELD and CMU-MOSEI datasets demonstrate that, although MRO activates fewer attention paths (as indicated by the mask), it maintains—or even improves—performance relative to full CMT. This selective routing mechanism leads to reduced computational cost and lightweight architecture, without compromising accuracy.

Looking ahead, MRO's dynamic routing and resource-efficient design provide a promising foundation for expanding to high-dimensional, heterogeneous domains, such as medical imaging (e.g., MRI–CT fusion), sensor-based prognostics (PHM), and satellite data fusion. To evaluate MRO's scalability beyond conventional tri-modal settings, we plan to test it on publicly available four-modal or higher datasets, such as MM-IMU (text, image, inertial, and audio) and MedFuse (CT, MRI, clinical notes, and genomics), with limited computational resources. Specifically, we aim to quantify how the number of activated paths, training efficiency, and modality-wise routing patterns evolve as the number of modalities increases. Although full-scale validation is reserved for future research, our architecture's modular design and prior-guided masking mechanism are inherently compatible with k-modal scaling ($O(k^2)$ path space) and enable graceful computational degradation through early-stage pruning. Recent work, such as HEALNet[10] empirically validates the feasibility and clinical value of four or more modalities in healthcare scenarios, which further supports the applicability of our framework. These planned extensions will demonstrate MRO's generalization capability in real-world multimodal environments and reinforce its potential as a versatile, efficient fusion strategy.

Acknowledgement

This research was financially supported by Hansung University.

References

[1] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6558–6569, 2019. DOI: https://doi.org/10.18653/v1/p19-1656

[2] B. Zoph and Q. V. Le, "Neural Architecture Search with Reinforcement Learning," *arXiv preprint arXiv:1611.01578*, 2016. DOI: https://doi.org/10.48550/arXiv.1611.01578

[3] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, "Once-for-All: Train One Network and Specialize it for Efficient Deployment," *Proc. Int. Conf. on Learning Representations (ICLR)*, 2020. DOI: https://doi.org/10.48550/arXiv.1908.09791

[4] P. Dong, X. Niu, L. Li, L. Xie, W. Zou, T. Ye, Z. Wei, and H. Pan, "Prior-Guided One-shot Neural Architecture Search," *Proc. AAAI Conf. on Artificial Intelligence*, vol. 38, no. 7, pp. 8947–8955, 2024. DOI: https://doi.org/10.48550/arXiv.2206.13329

[5] J. Wang, X. Wu, Y. Song, and S. Liu, "GraphCFC: An Efficient Graph-Based Method for Conditional Fusion Control," *IEEE Trans. Multimedia (TMM)*, 2023. DOI: https://doi.org/10.1109/TMM.2023.3260635

[6] M. Golovanevsky, E. Schiller, A. Nair, E. Han, R. Singh, and C. Eickhoff, "One-Versus-Others Attention: Scalable Multimodal Integration for Biomedical Data," *Proc. Int. Conf. on Artificial Intelligence in Medicine (AIME)*, 2024. DOI: https://doi.org/10.1142/9789819807024_0041

[7] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio, "Graph Attention Networks," *International Conference on Learning Representations (ICLR)*, 2018. DOI: https://doi.org/10.48550/arXiv.1710.10903

[8] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," *Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. DOI: https://doi.org/10.18653/v1/P18-1208

[9] S. Hazarika, G. Zimmermann, S. Poria, and R. Zimmermann, "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 527–536, 2019. DOI: https://doi.org/10.48550/arXiv.1810.02508

[10] K. Hemker, M. Jamnik, and N. Simidjievski, "HEALNet: Multimodal Fusion for Heterogeneous Biomedical Data," in *Proc. 38th Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2024. DOI: https://doi.org/10.48550/arXiv.2311.09115