Enhancing Multimodal Learning via Hierarchical Fusion Architecture Search With Inconsistency Mitigation

Kaifang Long, Guoyang Xie, *Member, IEEE*, Lianbo Ma, *Senior Member, IEEE*, Qing Li, *Senior Member, IEEE*, Min Huang, *Member, IEEE*, Jianhui Lv, and Zhichao Lu, *Member, IEEE*

***Abstract*—The design of effective multimodal feature fusion strategies is the key task for multimodal learning, which often requires huge computational costs with extensive expertise. In this paper, we seek to enhance multimodal learning via hierarchical fusion architecture search with inconsistency mitigation. Different from previous works, our Hierarchical Fusion Multimodal Neural Architecture Search (HF-MNAS) considers the inconsistency in modalities and labels, and fine-grained exploitation in multi-level fusion architectures. Specifically, it disentangles the hierarchical fusion problem into two-level (macro- and micro-level) search spaces. In the macro-level search space, the high-level and low-level features are extracted and then connected in a fine-grained way, where the inconsistency mitigation module is designed to minimize discrepancies between modalities and labels in cell outputs. In the micro-level search space, we find that different intermediate nodes in the cells exhibit different importance degrees. Then, we propose an importance-based node selection mechanism to form the optimal cells for feature fusion. We evaluate HF-MNAS on a series of multimodal classification tasks. Empirical evidence shows that HF-MNAS achieves competitive trade-off performance across accuracy, search time, and inference speed. In particular, HF-MNAS consumes minimal computational cost compared with state-of-the-art MNASs. Furthermore, we theoretically and experimentally verify that the modality-label inconsistency deteriorates the overall fusion performance of models such as accuracy and F1 score, and demonstrate that the proposed inconsistency mitigation module could effectively mitigate this phenomenon.**

Received 9 October 2024; revised 19 June 2025; accepted 12 August 2025. Date of publication 22 August 2025; date of current version 28 August 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62472079 and in part by the Fundamental Research Funds for the Central Universities under Grant N2417003. The associate editor coordinating the review of this article and approving it for publication was Prof. Jungong Han. *(Corresponding author: Lianbo Ma.)*

Kaifang Long and Lianbo Ma are with the College of Software, Northeastern University, Shenyang 110819, China (e-mail: kaifanglong@163.com; malb@swc.neu.edu.cn).

Guoyang Xie is with the Department of Intelligent Manufacturing, CATL, Ningde 201306, China, and also with the Department of Computer Science, City University of Hong Kong, Hong Kong, China (e-mail: guoyang.xie@ieee.org).

Qing Li is with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: liq@pcl.ac.cn).

Min Huang is with the College of Information Science and Engineering, State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China (e-mail: mhuang@mail.neu.edu.cn).

Jianhui Lv is with First Affiliated Hospital, Jinzhou Medical University, Jinzhou 121001, China (e-mail: lvjianhui2012@163.com).

Zhichao Lu is with the Department of Computer Science, City University of Hong Kong, Hong Kong, China (e-mail: luzhichaocn@gmail.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2025.3599673, provided by the authors.

Digital Object Identifier 10.1109/TIP.2025.3599673

***Index Terms*—Multimodal fusion, differentiable architecture search, attention mechanism.**

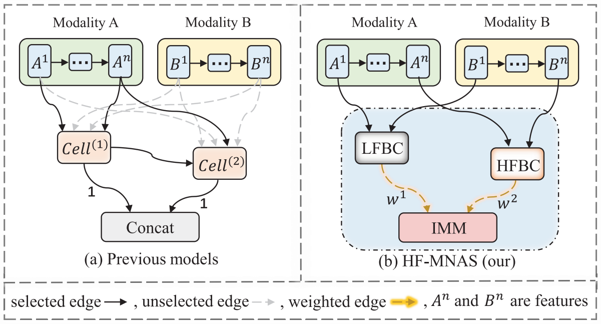
I. Introduction

With the explosive growth in advanced multimodal learning applications (e.g., action recognition [1], and image/video captioning [2]), how to obtain optimal multimodal feature fusion strategies becomes a major challenge that needs to be urgently solved. Conventional multimodal fusion approaches focus on the combination of multimodal feature vectors. However, their performance heavily relies on neural architectures, which are typically manual designs with extensive expertise. Recently, multimodal neural architecture search (MNAS) has emerged as a promising automatic design technique, aiming to search for optimal multimodal neural network models in an efficient way [3]. Instead of designing hand-crafted multimodal learning models based on extensive human expertise, MNAS is able to not only obtain competitive multimodal models as human experts do but also find new state-of-the-art fusion strategies [4].

However, existing MNAS methods suffer from the following two limitations:

❶ The search process of the feature fusion module is typically treated as a multimodal-coupled single-level search process, which maps all possible feature connection (fusion) combinations of all potential modalities into a uniform search space, and then searches the optimal connections from the vast search space, as shown in Fig. 1 (a). However, such a way is rather coarse-grained and neglects the discriminative functions of high-level and low-level features. As a result, many ineffective feature connections, e.g., unimodal connections, increase the redundancy of target multimodal search space and thus decrease the fusion performance, e.g., search efficiency and model accuracy. For example, Fig. 1 (a) shows the unsatisfactory multimodal fusion result found from the single-level search space, i.e., only a single modality is involved in the final fusion strategy.

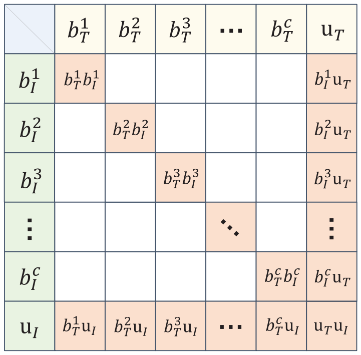
❷ The search of existing methods is based on an implicit assumption that all the intermediate nodes in each cell have the same contribution/importance over the performance of feature fusion, as shown in Fig. 1 (a). Little work focuses on the validity of the above assumption of MNAS. However, we discover that each intermediate node in the cell exhibits a specific contribution to the feature fusion. In this sense, as shown in Fig. 8, it is desired to apply importance-based



(a) Previous models

(b) HF-MNAS (our)

Fig. 1. Comparison of HF-MNAS with existing MNAS methods. In HF-MNAS, we re-design the macro-search space (blue dashed frame) and micro-search space (LFBC and HFBC) for multimodal fusion, in which LFBC and HFBC denote the low-level and high-level feature blend cell, respectively.



	b_T^1	b_T^2	b_T^3	⋯	b_T^c	u_T
b_I^1	$b_T^1 b_I^1$					$b_I^1 u_T$
b_I^2		$b_T^2 b_I^2$				$b_I^2 u_T$
b_I^3			$b_T^3 b_I^3$			$b_I^3 u_T$
⋮				⋱		⋮
b_I^c					$b_T^c b_I^c$	$b_I^c u_T$
u_I	$b_T^1 u_I$	$b_T^2 u_I$	$b_T^3 u_I$	⋯	$b_T^c u_I$	$u_T u_I$

Fig. 2. The combination rule of beliefs. Given the beliefs of the text (yellow block) and the image (green block), we recombine them to get new beliefs (orange block). The white block is the measure of belief conflict between the text and the image.

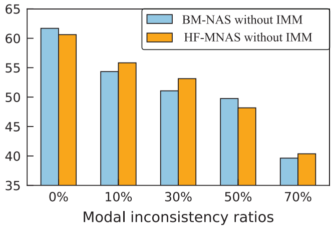


Fig. 3. Impact of the modalites inconsistency on F1-M score in MM-IMDB datasets.

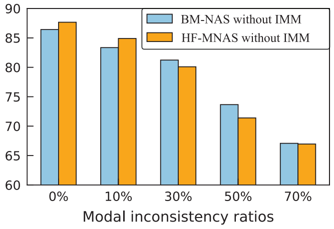


Fig. 4. Impact of the modalites inconsistency on F1-M score in HARM_P datasets.

node selection to form the final cells of feature fusion. Moreover, modalities inconsistency (also known as modality-label inconsistency) between cell outputs also influences the overall fusion performance, which is not considered in existing MNAS works.

In this research, we provide a solution called HF-MNAS (Hierarchical Fusion Multimodal Neural Architecture Search) to overcome these restrictions. HF-MNAS decouples the issue into two levels, namely macro and micro, inside the multi-modal search spaces. Considering the constraint ❶, we introduce a new macro-level search space, as seen in Fig. 1 (b). HF-MNAS utilizes a hierarchical fusion approach to combine different dimension modal features obtained by the unimodal feature extraction module (UFEM). This approach involves a low-level feature blend cell (LFBC) and a high-level feature blend cell (HFBC). The purpose of this fusion is to improve search speed and address the issue of underutilization of modal features. In addition, to tackle the issue of inconsistency interference between modalities and labels (Limitation ❷), we employ an inconsistency mitigation module (IMM) that comprises deep canonical correlation analysis (DCCA) and multi-head attention. This module aims to enhance the correlation between the LFBC and the HFBC, allowing the intermediate nodes to effectively capture shared features across different modalities during fusion. Within the micro-level search space, we observe that various intermediate nodes in the cell possess varying degrees of significance. Consequently, we suggest a technique for selecting nodes based on their relevance in order to create an ideal cell for feature fusion. Empirical evidence demonstrates that HF-MNAS is superior to other MNAS. In detail, HF-MNAS improves 1.60% and 0.98% in F1-W and F1-M on the MM-IMDB dataset. On the HARM_P dataset, HF-MNAS boosts 5.32% and 3.37% in F1-M and accuracy. On the HARM_C dataset, HF-MNAS improves 4.03% and 2.01% in F1-M and accuracy.

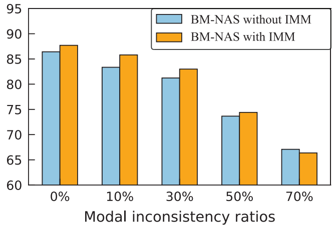


Fig. 5. The F1-M scores of the BM-NAS method with the IMM module on HARM_P.

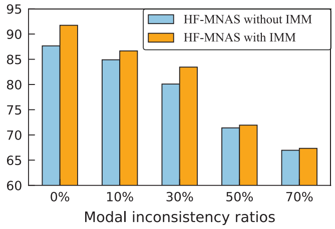


Fig. 6. The F1-M scores of the HF-MNAS method with the IMM module on HARM_P.

The contributions of this paper are as follows:

- We present an efficient multimodal fusion design via hierarchical fusion architecture search with inconsistency mitigation. Our architecture search framework separates the fusion search space into macro and micro levels, where the IMM module is specifically utilized to remove the inconsistency in modalities and labels in cell outputs.
- We provide a theoretical analysis from the mathematical perspective, revealing that the modality-label inconsis-

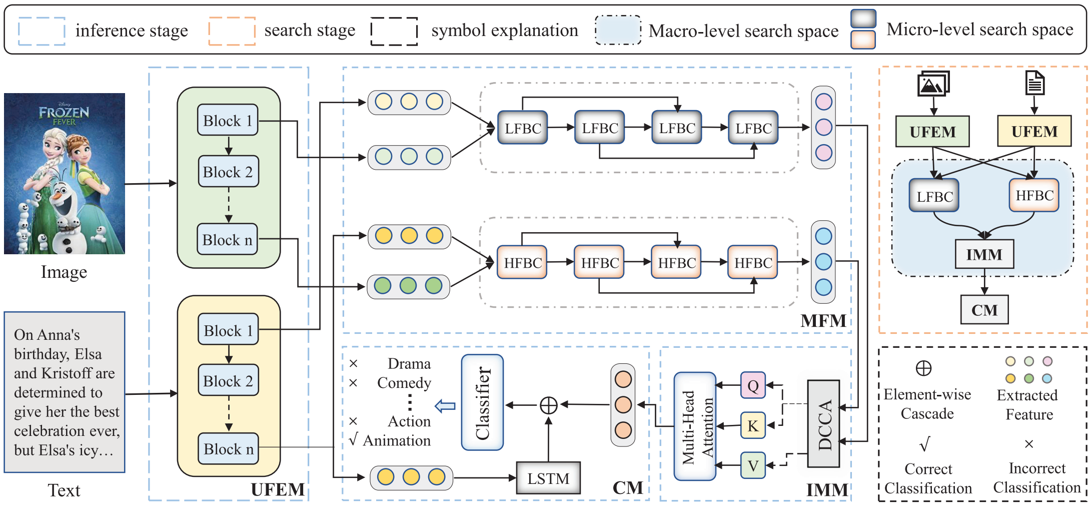


Fig. 7. The network architecture of HF-MNAS. The micro-level search space consists of low-level feature blend cell (LFBC) and high-level feature blend cell (HFBC), which are responsible for hierarchical fusion. The macro-level search space consists of multimodal fusion module (MFM) and inconsistency mitigation module (IMM), where UFEM denotes unimodal feature extraction module, and CM is the classification module.

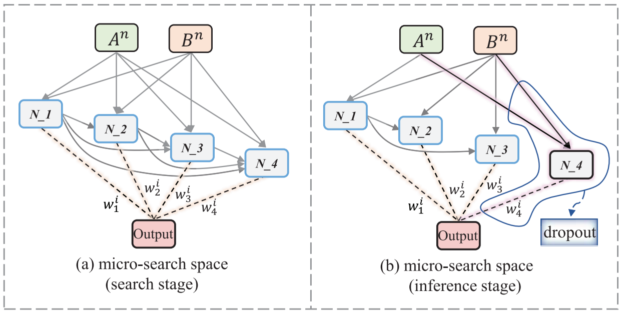


Fig. 8. The micro-search spaces (LFBC or HFBC). *N* denotes the intermediate node. *w* means the node weight. Dropout denotes the node output is not used in the inference phase.

tency could worsen the overall fusion performance of models such as accuracy and F1 score. Moreover, the effectiveness of our proposed IMM module for alleviating the above phenomena is also verified on a set of multimodal fusion methods.

- For the design of micro search space, we observe varying degrees of significance among distinct intermediate nodes in a cell. Depending on this observation, we propose a technique for selecting nodes depending on their importance to create an ideal cell for feature fusion.

Outline. The other parts of this paper are as follows. In Section II, we summarize related work on multimodal fusion and multimodal neural architecture search. In Section III, We theoretically and experimentally investigate the phenomenon that the modalities' inconsistency deteriorates the overall fusion performance of models. In Section IV, we illustrate our methodology in detail. In Section V, we perform extensive experiments to validate our proposed model and deeply analyze the results. Lastly, we conclude this paper in Section VI.

II. Related Work

A. Multimodel Fusion

Multimodal fusion has been widely integrated with various deep neural networks to enhance performance by exploiting multiple modalities (e.g., text, image, and audio) simultaneously [5], [6], [7], [8], [9], and it can be divided into two categories: early fusion which performs combination at the input level and late fusion which performs combination at the decision level. Accordingly, many efforts have pushed multimodal fusion forward. For example, [10] proposed gated multimodal units to determine the impact of different modality features on hidden units. Reference [11] achieved excellent performance on the text-visual sentiment analysis task using a gated multimodal embedding layer and an LSTM layer with temporal attention. Reference [12] proposed an emotion classification method based on multimodal signals using the synergistic effect of multiple neural networks.

However, early fusion fails to consider the complementary information owing to excessive redundancy features between modalities, while late fusion misses the better features in feature extraction process since it performs fusion at the decision level. Thus, the hybrid fusion approach emerges as an effective way. For example, [1] enhanced the effectiveness of modal fusion by integrating information from multiple media to generate optimal decisions. Reference [13] utilized a co-attention mechanism to interact with features. Reference [14] proposed a multimodal transfer module that can be embedded in various layers to realize multimodal fusion. Reference [15] proposed a bi-directional transformer structure for text-image classification. In essence, these works resort to enhanced networks to build a good modal fusion model. However, these methods are built upon existing network architectures, limiting the scope of architecture exploration for better fusion.

B. Multimodel Neural Architecture Search

Starting from Google Brain's work [16], a series of neural architecture search (NAS) methods have been proposed and developed to jointly optimize the network architecture and weights for various learning tasks [17], [18], [19], [20]. Popular search paradigms to realize these NAS methods include Bayesian optimization [21], reinforcement learning [22], [23], and evolutionary algorithms [24], [25], [26], [27]. However, the computational burden of these types of methods is still unaffordable. Thus, [28] proposed an intuitive solution, Differentiable Architecture Search (DARTS), which introduces a continuous and relaxation strategy to make the gradient-based search paradigm feasible, thus dramatically boosting the search process.

With the rapid development of NAS, it has shown tremendous potential in multimodal learning [29], [30], [31] since it is able to search for the best multimodal network models automatically, instead of human expertise to design manually. For example, [3] proposed an unimodal features search strategy using the SMBO algorithm. Reference [30] proposed a generic MMnas framework that exploits NAS to find the optimal architecture for different tasks. Reference [32] proposed a novel search space and used evolutionary NAS to obtain optimal models for electronic health record tasks. Recently, a bi-level multimodal neural architecture search framework (BM-NAS) was proposed and obtained satisfactory performance [4]. Specifically, in the upper level, BM-NAS selects the input pairs for cells from the pre-trained unimodal backbone model, and in the lower layer, it selects the fusion operation for each intermediate node from the search space. In fact, BM-NAS maps all possible combinations of feature connections into a uniform search space during the feature selection process, which ignores the discriminative function between high- and low-level features. Different from BM-NAS, our HF-MNAS decouples the multimodal coupled single-level search process into a macro- and micro-level multimodal search spaces, where the identification and fusion of low- and high-level features are considered at macro space. Moreover, HF-MNAS considers the negative impact of the modalities' inconsistency from cell outputs, and handles it with inconsistency mitigation operation.

III. Modalities' Inconsistency Issue

In this section, we conduct theoretical analyses and empirical studies to verify the negative impact of the modalities' inconsistency over the overall fusion performance.

A. Problem Statement

In the multimodal fusion tasks, modalities' inconsistency, also known as modality-label inconsistency, refers to multimodal data exhibiting differences in representing label information [33], [34], [35], [36]. Specifically, suppose that there exist two modalities, i.e., text (T) and image (I). For the c-class classification task, we obtain a set of category probabilities $e_T = [e_t^1, \ldots, e_t^j, \ldots, e_t^c]$ and $e_I = [e_i^1, \ldots, e_i^j, \ldots, e_i^c]$ through the target model for T and I, where $e_t^j \in e_T$ and $e_i^j \in e_I$ denote the j-th category probability in T and I, respectively, and j represents the ground-truth label. If the two modalities can consistently characterize label information, then e_t^j should be maximum in e_T, and e_i^j should be maximum in e_I. If the modalities and labels information are inconsistent, it is possible that e_t^j or e_i^j is not maximum in e_T or e_I, and may even be minimum.

If we integrate text and images that are modalities and labels inconsistent via the fusion operator, the probability of belonging to the ground-truth label is potentially decreased when using the fused features for classification. Therefore, how to mitigate the overall fusion performance degradation issue arising from the modalities' inconsistency is crucial for multimodal fusion.

B. Theoretical Analysis

Motivated by the principle of Dempster-Shafer evidence theory (DST) [37], we observe that belief mass (b) and uncertainty (u) are pivotal to estimating the trustworthiness of the target model's prediction results [38], [39]. For this perspective, we can utilize the DST principle to scrutinize mechanisms for the impact of modalities' inconsistency on the overall performance of multimodal fusion architectures.

For the multimodal c-class classification task with text-image pairs, suppose that the category probabilities for each modality have been acquired, e.g., the text modality is $e_T = [e_t^1, e_t^2, e_t^3, \ldots, e_t^c]$ and the image modality is $e_I = [e_i^1, e_i^2, e_i^3, \ldots, e_i^c]$. Then, we apply DST to obtain each modality's opinion, each of which consists of uncertainty and belief mass, e.g., $\{u_T, \{b_T^j\}_{j=1}^c\}$ for text modality opinion and $\{u_I, \{b_I^j\}_{j=1}^c\}$ for image modality opinion. For the text modality opinion, its uncertainty u_T and belief mass b_T^j can be acquired by:

$$b_T^j = \frac{e_t^j}{S_T}, \quad u_T = \frac{c}{S_T}, \tag{1}$$

$$u_T + b_T^1 + b_T^2 + \ldots + b_T^c = 1, \tag{2}$$

$$S_T = (e_t^1 + 1) + (e_t^2 + 1) + \ldots + (e_t^c + 1), \tag{3}$$

where b_T^j is the belief mass of the j-th class of the text modality, and the probability of belonging to the j-th class increases as b_T^j grows. u_T indicates the trustworthiness of the classification, and the smaller u_T means the better classification accuracy. For the image modality, the corresponding u_I and b_I^j are obtained similarly to the case of text modality.

Given the opinions derived from the text (i.e., $\{u_T, \{b_T^j\}_{j=1}^c\}$) and image (i.e., $\{u_I, \{b_I^j\}_{j=1}^c\}$) modalities, we leverage the combination principle of DST to theoretically analyze whether integrating text and image modality to form a new opinion (i.e., $\{u_{fuison}, \{b_{fuison}^j\}_{j=1}^c\}$) influences the model's classification accuracy. Here, b_{fuison}^j and u_{fuison} indicate new belief mass and uncertainty, as following:

$$b_{fusion}^j = \frac{1}{1-K}(b_T^j b_I^j + b_T^j u_I + b_I^j u_T), \tag{4}$$

$$u_{fusion} = \frac{1}{1-K}(u_I u_T), \tag{5}$$

where $\frac{1}{1-K}$ denotes the normalized scaling factor, and $K = \sum_{i \neq j} b_T^i b_T^j$ is a measure of the conflict amount between text and image belief mass, as shown in the white block of Fig. 2.

Therefore, we derive the following propositions:

Proposition 1: Suppose that the modalities are consistent, in the case $b_I^g \geq b_T^{max}$, where b_I^g is the belief mass of the index g of ground-truth label in image, and b_T^{max} is the maximum belief mass in text, integrating the image modality in the text modality makes the new belief mass satisfy $b_{fusion}^g \geq b_T^g$.

Proof 1.

$$\begin{aligned} b_{fusion}^g &= \frac{b_T^g b_I^g + b_T^g u_I + b_I^g u_T}{\sum_{c=1}^C b_T^c b_I^c + u_T + u_I - u_T u_I} \\ &\geq \frac{b_T^g (b_I^g + u_I + u_T)}{b_T^{max} + u_I + u_T} \geq b_T^g, \end{aligned}$$

Proposition 2: When text and image modalities are divergent in subjective opinions or semantic representations (i.e., the modalities' inconsistency), the new overall uncertainty u_{fusion} increases accordingly. That is, u_{fusion} is positively correlated with u_T and u_I.

Proof 2.

$$\begin{aligned} u_{fusion} &= \frac{u_T u_I}{\sum_{c=1}^C (b_I^c u_T + b_T^c u_I + b_I^c b_T^c) + u_T u_I} \\ &= \frac{1}{\sum_{c=1}^C \left(\frac{b_T^c b_I^c}{u_T u_I} + \frac{b_I^c}{u_I} + \frac{b_T^c}{u_T} \right) + 1}. \end{aligned}$$

Notably, more detailed proofs are provided in Appendix (see Supplementary Material).

From the aforementioned propositions, we can obtain the following conclusions: (1) if the modalities and labels are consistent, we can have the belief masses b_T^g and b_I^g in both text and image modalities should be larger than the corresponding u_T and u_I, respectively. According to ***Proposition*** 1, we can get $b_{fusion}^g \geq b_T^g$. That is, integrating image modality into text modality boosts the accuracy of model classification. (2) If modalities and labels are inconsistent, it may imply that only one of the modalities in the text and image reflects the ground-truth label (i.e., u_T or u_I increases) or both modalities fail to reflect the ground-truth label (both u_T and u_I increase). According to ***Proposition*** 2, where the u_{fusion} is positively related to the overall uncertainty (u_T or u_I) of each modality, we can have u_{fusion} becomes larger as u_T or u_I increases, degrading the model classification accuracy.

C. Experimentation on the Modalities' Inconsistency Issue

To empirically verify that modalities inconsistency significantly deteriorates model performance, necessitating adjusting the inconsistency of text-image pairs in MM-IMDB and HARM_P datasets. For MM-IMDB, we first partition it into genre-based subsets according to the ground-truth labels of image-text pairs. Then, we select an original sample from a subset and randomly choose a substitute sample sharing at least one same genre label from the same subset. Finally, replacing the original text with the substitute sample text to generate inconsistent data with semantic misalignment. Regarding HARM_P, we divide it into three subsets according to the not harmful, somewhat harmful, and very harmful labels, before generating inconsistent pairs by selecting original samples from very harmful or somewhat harmful subsets and replacing their text from the not harmful subset.

As shown in the horizontal coordinate of Figs. 3–6, 0% represents the original MM-IMDB or HARM_P dataset, 10% indicates that 10% of the text-image pairs in the dataset are inconsistent, 30% denotes 30% inconsistent text-image pairs in the dataset, and so forth. Subsequently, we analyze the experimental results in detail. As shown in Fig. 3, the macro F1 scores of both BM-NAS [4] and HF-MNAS (without inconsistency mitigation module) exhibit a dramatic performance degradation on the MM-IMDB dataset as the modal inconsistency rate increases. Similarly, Fig. 4 reveals a consistent performance decline for these methods on the HARM_P dataset, clearly demonstrating that modality inconsistency severely impairs the overall fusion performance of the models. To mitigate this negative effect, we propose an inconsistency mitigation module (more details are given in Section IV-C). As shown in Figs. 5 and 6, we embed the inconsistency mitigation module into BM-NAS and HF-MNAS, and find that both of them outperform their original versions to some extent on HARM_P dataset. These encouraging results validate the effectiveness of our proposed inconsistency mitigation module.

IV. Methodology

As shown in Fig. 7, we propose HF-MNAS, which disentangles the NAS problem into macro-level and micro-level multimodal search spaces. In the following, we describe the unimodal feature extraction module in Section IV-A, the micro-level search space in Section IV-B, the macro-level search space in Section IV-C, and the architecture search and evaluation in Section IV-D.

A. Unimodal Feature Extraction Module

Following previous multimodal fusion approaches [1], [3], [4], HF-MNAS uses the same pre-trained unimodal backbone model as the feature extractor for fair comparison. Specifically, we employ Maxout MLP as the backbone model for extracting text modalities (T_{low} and T_{high}) and VGG Transfer as the backbone model for extracting image modalities (I_{low} and I_{high}). The formulas are as follows:

$$I_{low} = VGG_{net\_1}(X_I), \tag{6}$$

$$I_{high} = VGG_{net\_n}(X_I), \tag{7}$$

$$T_{low} = Maxout\_MLP_{net\_1}(X_T), \tag{8}$$

$$T_{high} = Maxout\_MLP_{net\_n}(X_T), \tag{9}$$

where X_I and X_T denote the inputs to the unimodal backbone model, i.e., image modality and text modality. I_{low} and I_{high} are the lower-level and the higher-level image features that are obtained by using VGG Transfer with different numbers of blocks. T_{low} and T_{high} are the lower-level text features and the higher-level text features which are obtained by using Maxout MLP with different numbers of blocks.

B. The Micro-Level Search Space

After obtaining low-level and high-level text/image features, we elaborate the micro-level search space based on DATRS [28], i.e., the operation search space of LFBC or HFBC on edges and intermediate nodes. As shown in Fig. 8-(a), the

LFBC and HFBC are directed acyclic graphs consisting of N nodes (two input nodes, one output node, and $N-3$ intermediate nodes) and {(N-2)!-1} edges, where each intermediate node is connected to all its predecessor nodes via a directed edge. It is worth noting that the input of LFBC is (T_{low}, I_{low}) and the input of HFBC is (T_{high}, I_{high}).

1) Operation Search Space on Edge: In our work, each directed edge contains ten different types of primitive search operations, which are 3×3 max pooling, 3×3 average pooling, 3 × 3 separable convolution, 3 × 3 dilated convolution, skip connection, none operation, and Linear transformations with four different activation functions (ReLU, Sigmoid, Tan, and ELU) [4], [40].

After determining the search space of edge operations, as shown in Eq. 10, we get the influence of predecessor nodes and their corresponding edges towards the current node using the continuous concept of DARTS. That is, we obtain the weights of 10 different types of operations in each directed edge. Then, we use the relaxation concept to discrete the operations, where each intermediate node is connected to two predecessor nodes by the two operation edges with the largest weights. For more details, please refer to [28].

$$\tilde{o}p^{(i,j)}(x_i) = \sum_{op\in O} \frac{exp\{\gamma_{op}^{(i,j)}\}}{\sum_{op'\in O} exp\{\gamma_{op'}^{(i,j)}\}} \cdot op(x_i), \tag{10}$$

where O denotes the operation search space on edges. $\gamma_{op}^{(i,j)}$ indicates the continuous coefficient. x_i represents the feature matrix of the i-th node, and $op(\cdot)$ denotes to operate on x_i.

2) Operation Search Space on the Intermediate Node: After determining the predecessor nodes and their corresponding edges for each intermediate node, we additionally assign an operations search space to the intermediate nodes. Our goal is to find the best fusion operation for the two input feature matrices of the current node. The equations are as follows:

$$\tilde{f}^{(node)}(x_1, x_2) = \sum_{f\in F} \frac{exp\{\beta_f^{(node)}\}}{\sum_{f'\in F} exp\{\beta_{f'}^{node}\}} \cdot f(x_1, x_2), \tag{11}$$

$$f^{(node)} = \arg\max_{f\in F} \ \beta_f^{(node)}, \tag{12}$$

where x_1 and x_2 are the feature matrices. f represents the fusion operation of the feature matrix in the node. β indicates the weight of the fusion operation. F denotes the operation search space of the intermediate nodes, which are as follows:

$Sum(x_1, x_2)$: We fuse the feature matrices from different modalities by the sum operation.

$$Sum(x_1, x_2) = x_1 + x_2. \tag{13}$$

$MHAtt(x_1, x_2)$: The multi-head attention mechanism can improve the model's feature extraction capability by computing attention weights through several independent attention heads, each capturing distinct feature relationships. The equations are as follows:

$$head_j^1 = softmax\left(\frac{Q_1 K_2^T}{\sqrt{d_{K_2}}}\right) \cdot V_1, \tag{14}$$

$$head_j^2 = softmax\left(\frac{Q_2 K_1^T}{\sqrt{d_{K_1}}}\right) \cdot V_2, \tag{15}$$

$$\begin{aligned} MHAtt(x_1, x_2) = Concat(&head_1^1, head_2^1, \\ &\ldots, head_h^1, head_1^2, head_2^2, \ldots head_h^2) \cdot W_o, \end{aligned} \tag{16}$$

where $Q_{1/2} = W_{Q_{1/2}} \cdot x_{1/2}$ denotes the query matrix, $K_{1/2} = W_{K_{1/2}} \cdot x_{1/2}$ represents the key matrix, and $V_{1/2} = W_{V_{1/2}} \cdot x_{1/2}$ corresponds to the value matrix. $\sqrt{d_k}$ is the dimension of the matrix. Furthermore, Scaled dot product attention (i.e., $ScaleDotAtt(x_1, x_2)$) and Bidirectional attention (i.e., $Bidirectional\_Attention(x_1, x_2)$) are two potential fusion operations, which are also integrated into the operation search space for seeking better fusion architecture.

$Squeeze\_Excitation(x_1, x_2)$: This makes the network pay more attention to the meaningful features for the classification task by learning the importance weights between the feature channels. As follows, S_{x_1} means compressing the feature vector x_1 by global average pooling to obtain the global features for each channel. E_{x_1} refers to enhancing the response for important features by the activation function.

$$SE(x_1, x_2) = E_{x_1} \cdot x_2, \tag{17}$$

$$E_{x_1} = \sigma(S_{x_1} \cdot W + b) \cdot x_2, \tag{18}$$

$$S_{x_1} = \frac{1}{L} \sum_{i=1}^{L} x_1(B, C, i). \tag{19}$$

$LinearGLU(x_1, x_2)$: This operation transforms the features x_1 and x_2 by using a gated linear unit to facilitate the contribution of different modalities to fusion.

$$\begin{aligned} LinearGLU(x_1, x_2) &= GLU(x_1 W_1, x_2 W_2) \\ &= x_1 W_1 \odot sigmoid(x_2 W_2). \end{aligned} \tag{20}$$

$ConcatFC(x_1, x_2)$: This operation means that two different modalities are cascaded, and the linear layer with the ReLU activation function is used to make a linear transformation.

$$ConcatFC(x_1, x_2) = ReLU(Concat(x_1, x_2)W + B). \tag{21}$$

$Multiply(x_1, x_2)$: We fuse the feature matrices from different modalities by the element-wise multiplication operation.

$$Multiply(x_1, x_2) = x_1 \cdot x_2. \tag{22}$$

$Mamba\_fusion(x_1, x_2)$: This operation [41] has potential strengths for extracting fine-grained multimodal features and efficiently modeling cross-modal correlations. As follows, Dwc (·) indicates depthwise convolution operation, ES2D (·) denotes efficient spatial scanning 2D operation and SiLU (·) represents the activation function.

$$\begin{aligned} &\tilde{M} = Dwc(linear(x_1)) \cdot Dwc(linear(x_2)), \\ &m_{1/2} = ES2D(SiLU(\tilde{M})) \cdot SiLU(linear(x_{1/2}), \\ &Mamba\_fuison(x_1, x_2) = m_1 + m_2 \end{aligned} \tag{23}$$

C. The Macro-Level Search Space

As shown in Fig. 7, our macro-level search space consists of MFM which uses LFBC and HFBC to perform feature fusion, and NEM to alleviate modal inconsistency. The main processes and mechanism are as follows.

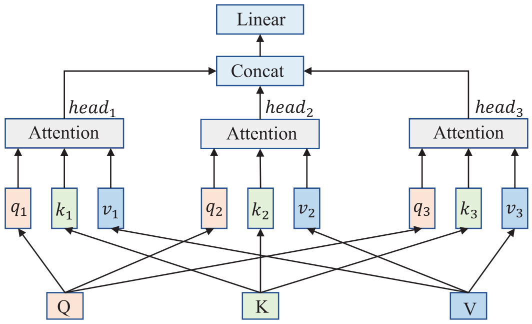


Fig. 9. Example of a multi-head attention mechanism. Q, K, and V represent the query matrix, key matrix, and value matrix respectively.

1) Multimodal Fusion Module: In this module, we use a hierarchical fusion approach to combine different dimension modal features obtained by UFEM. This approach involves LFBC and HFBC. For low-level text and image features (T_{low}, I_{low}), we perform fusion using LFBC. For high-level text and image features (T_{high}, I_{high}), we use HFBC for fusion. Furthermore, we observe that different intermediate nodes in the cells exhibit different importance degrees. Based on this, we propose an importance-based node selection mechanism to form the optimal cells. As shown in Fig. 8, we assign a weight to each intermediate node and remove the node's output with less influence in the inference phase based on the weight. The formulas are given below:

$$\begin{aligned} C_1 &= LFBC\,(T_{low}, I_{low}) \\ &= W_{c1}^1 N_{c1}^1 \oplus W_{c1}^2 N_{c1}^2 \oplus \cdots \oplus W_{c1}^n N_{c1}^n, \end{aligned} \tag{24}$$

$$\begin{aligned} C_2 &= HFBC\,(T_{high}, I_{high}) \\ &= W_{c2}^1 N_{c2}^1 \oplus W_{c2}^2 N_{c2}^2 \oplus \cdots \oplus W_{c2}^n N_{c2}^n, \end{aligned} \tag{25}$$

where C_1 and C_2 denote the output of LFBC and HFBC, respectively, N denotes the intermediate node, $\oplus$ indicates cascade, W means the weight of the node importance, and HF-MNAS determines whether to retain the node output or not based on weights.

2) Inconsistency Mitigation Module: DCCA [42] is a statistical analysis technique that sufficiently combines neural networks with canonical correlation analysis to reflect the overall correlation between two groups of feature vectors. In our work, we utilize the method to optimize the correlation between different modal features. As shown in Eqs. 26- 27, f denotes the neural network. C denotes the output of LFBC or HFBC. w indicates the parameters. O represents the output. Shown in Eq. 28, the purpose of DCCA is to jointly learn w_1 and w_2 to make a high correlation between O_1 and O_2.

$$O_1 = f_1(C_1, w_1), \tag{26}$$

$$O_2 = f_2(C_2, w_2), \tag{27}$$

$$(w_1^*, w_2^*) = \underset{w_1, w_2}{\arg\max}\; corr(f_1(m_1, w_1), f_2(m_2, w_2)). \tag{28}$$

After obtaining the features optimized by DCCA, as shown in Fig. 9, we employ multi-head attention (MHA) [43] to strengthen the network's focus on common attributes between O_1 and O_2. Firstly, we compute the query (Q), key (K), and value (V) matrices for both O_1 and O_2. Subsequently, leveraging the capability of each attention head to capture distinct feature relationships, we aggregate the outputs from multiple heads to construct the final multi-head attention representation. It is noteworthy that the synergistic usage of DCCA and MHA achieves dual benefits: (1) effectively mitigating the interference caused by modality-label inconsistency, while (2) facilitating enhanced feature interaction between LFBC and HFBC to optimize multimodal fusion performance.

D. Architecture Search and Evaluation

1) Architecture Parameters: The weights of the primitive operations (γ) on the edges are shown in Eq. 10, the weights of the primitive operations (β) on the intermediate nodes are shown in Eq. 11, and the weights of the intermediate nodes importance (w^i) are shown in Fig. 8. γ is used for the operation selection of the features for the input cells (i.e., LFBC and HFBC). β is used for the operation selection of multimodal fusion within the intermediate nodes. w^i is used to rank the importance of intermediate nodes.

2) Search Algorithm: After constructing the neural network model through the defined micro-level search space and macro-level search space, we use DRATS [28] to alternately optimize the architecture parameters and network parameters, where the architecture parameter contains γ and β aiming to find the optimal structure, and the network parameter involves the parameter w^i of the intermediate node importance. The equations are shown as follows:

$$\min_{\alpha} \quad Loss_{val}(w^*(\gamma,\beta), (\gamma,\beta)), \tag{29}$$

$$s.t. \quad w^*(\gamma,\beta) = \underset{w}{\arg\min}\; Loss_{train}(w, (\gamma,\beta)), \tag{30}$$

where $Loss_{train}$ and $Loss_{val}$ represent the training and validation losses, respectively, which are jointly determined by the network parameter (w) and the architecture parameter (γ,β). The optimization of the network parameter and architecture parameter iterations are shown in the following steps:

❶ Firstly, We initiate the network parameter (w) and the architecture parameter (γ,β) of the model.

❷ Secondly, we minimize the training loss by fixing γ and β to obtain the network weights $w^*(\gamma,\beta)$.

❸ Then we minimize the validation loss by fixing $w^*(\gamma,\beta)$ to obtain the architecture parameter γ and β. Finally, the optimization of w and (γ,β) is iteratively performed to get better structure parameters. Specifically, in Algorithm 1, we describe the architecture search process of HF-MNAS.

Note that once the above iterations converge, we utilize the architecture parameter obtained from training to discretize the operations to obtain the optimal network structure and retrain the structure for the final evaluation.

3) Evaluation: In the architecture evaluation, we select the LFBC and HFBC structures with the best validation performance as our multimodal fusion module. In particular, since text features contain abundant semantic information, we use bidirectional long and short-term memory (Bi-LSTM) for processing in the classification module. Finally, the classification is performed by cascading it with IMM output.

In the inference phase, as shown in the blue dashed box in Fig. 8, we hierarchically stacked the LFBC and HFBC as the

Algorithm 1 The Architecture Search Process of HF-MNAS

```
Input: Multimodal datasets (text-image pairs), including training, validation and test data.
Output: The genotype of architecture networks.
Define the operation search space on edges and intermediate nodes.
Initialize architecture parameters and network parameters.
for each text-image pair ($X_T$ and $X_I$) do
    Extract image low-level and high-level features by Eqs. (6) and (7). $I_{low}, I_{high} \leftarrow VGG(X_I)$.
    Extract text low-level and high-level features by Eqs. (8) and (9). $T_{low}, T_{high} \leftarrow Maxout\_MLP\ (X_T)$.
end for
for $e \leftarrow 1$ to Epochs do
    Use $I_{low}$ and $T_{low}$ as $LFBC$ inputs. Use $I_{high}$ and $T_{high}$ as $HFBC$ inputs.
    Fixed network parameters to train the architecture parameters, i.e., to obtain the values of each operation weight on nodes and edges for $LFBC$ and $HFBC$.
    Fixed architectural parameters to train the network parameters, i.e., to obtain the importance degree of each intermediate node in $LFBC$ and $HFBC$.
    Using $DCCA$ and attention mechanism to fuse LFBC and HFBC outputs according to Eqs. (24) and (26).
    Obtain the final multimodal fusion features for classification.
end for
Return the final architecture genotype.
```

final multimodal fusion module. Then, we used the training set and validation set to jointly train the unimodal model and the searched structures.

V. Experiments

In this section, we conduct experiments on five publicly available multimodal datasets, which are MM-IMDB, HARM_P, HARM_C, NTU RGB-D, and EgoGesture. Compared with the state-of-the-art methods, our proposed approach performs better on different evaluation metrics.

A. Datasets and Settings

1) Datasets: The MM-IMDB dataset was proposed by Ovalle et al. [10]. The dataset contains 26 different film genres such as drama, comedy, horror, etc. and the film genres are classified based on the image posters and the plot of films. The HARM_P and HARM_C datasets were proposed by Pramanick et al. [44] and composed of harmful memes related to US politics and COVID-19, respectively, where the image modality is collected from Google Chrome and several social platforms such as Reddit, and the text modality consists of words extracted from the images. NTU RGB-D was proposed by Shahroudy et al. [45] as a large-scale dataset for multimodal human action recognition, with action sequences captured by both depth and RGB cameras, encompassing 56 diverse human action categories. EgoGesture was proposed by Zhang et al. [46] as a multimodal gesture recognition dataset consisting of RGB and depth images covering 50 different subjects and 6 distinct scenarios.

TABLE I
Statistics of Datasets, Where I, T, V, and P Represent Image, Text, Video, and Pose, and R and D Denote RGB and Depth Image, Respectively

Datasets	Modalities	*Samples*		
		Train	Dev	Test
MM-IMDB	I+T	15552	2608	7799
HARM_P	I+T	3020	177	355
HARM_C	I+T	3013	177	354
NTU RGB-D	V+P	23760	2519	16558
EgoGesture	R+D	14416	4768	4977

As shown in Table I, each of the five datasets is divided into a train set, a development set, and a test set. For a fair comparison, the division of the datasets used in our experiments is consistent with baselines. For example, the MM-IMDB dataset is divided into 60% as a train set, 10% as a development set and 30% as a test set, and the HARM_P and HARM_C datasets are divided into 85% as a train set, 5% as a development set and 10% as a test set, the NTU RGB-D/EgoGesture dataset is divided into 55/40% as a train set, 5/20% as a development set and 40/20% as a test set.

2) Implementation Details: Our method adopts a single A100 GPU (80GB) for training, with the batch size set to 8, epoch set to 30, and dropout set to 0.1. The learning rate and weight decay rate for the architecture parameters are set to $3e^{-4}$ and $1e^{-3}$, respectively. The max learning rate, min learning rate, and weight decay rate for the network parameters are set to $1e^{-3}$, $1e^{-6}$, and $4e^{-3}$, respectively. Furthermore, we use Adma to optimize the network and architecture parameters.

3) Evaluation Metrics: We assessed our model by utilizing widely used evaluation metrics. Specifically, for the MM-IMDB dataset, we used the weighted F1 score and macro F1 score to evaluate the effectiveness of HF-MNAS. For the HARM_P and HARM_C datasets, we use the accuracy and the macro F1 score to evaluate the effectiveness of HF-MNAS. For the NTU RGB-D and EgoGesture datasets, we employ accuracy to evaluate the effectiveness of HF-MNAS.

B. Baselines

For performance comparison, we use the following baselines:

Unimodal baselines

- **Maxout_MLP** [47], maxout is a function approximator. A standard multilayer perceptron (MLP) network can approximate arbitrary functions as a way to improve task performance if the hidden layer contains enough neurons.
- **BERT** [48], a pre-trained language model, excels at capturing textual semantics and contextual knowledge.
- **BERT+Prompt** [49], Petroni et al. further improved BERT to enhance its ability to model sequences.
- **VGG** [50] attempts to build a deep network by using small convolutional kernels to explore the importance of the network's depth on image recognition accuracy.

- **DenseNet-161** [51] is a dense convolutional network which proves that convolutional networks can be trained more efficiently where there are connections between layers near the input and close to the output.
- **ResNet-152**, He et al. [52] proposed a residual learning framework to simplify the training of networks that are deeper than previously used networks.
- **ResNeXt-101**, Xie et al. [53] proposed a highly modular network architecture for image classification by repeating a building block.
- **Inflated ResNet-50**, Baradel et al. [54] used the recurrent spatial attention model to process features extracted from different local glimpses for action recognition.
- **Co-occurence**, Li et al. [55] proposed an end-to-end convolutional co-occurrence feature learning framework for human action recognition.

Multimodal baselines

- **Two-stream** [56] is a two-stream ConvNet architecture containing spatial and temporal networks that efficiently improves the performance of the model.
- **GMU** [10] is a gated neural network model for multimodal learning that ensembles multimodal representations by fusing features from different modalities.
- **CentralNet** [1] is a multimodal fusion method that generates an optimal strategy by aggregating information from diverse multimodalities.
- **MFAS** [3] utilizes NAS to address the multimodal classification problem, which aims to find the best architecture for a given dataset using search space operations.
- **MMBT** [15] is a supervised multimodal bidirectional transformer that fine-tunes an unimodal encoder by combining image and text information.
- **BM-NAS** [4] is a bilayer multimodal NAS framework that allows efficient search of unimodal features and multimodal features for fusion operations.
- **MMBV** [57] is a Multimodal BERT-ViT model that improves task performance by focusing on the use of weaker modal information and regularized loss function.
- **ViLBERT**, Lu et al. [13] proposed a visual language BERT to learn image features and text semantics.
- **VisualBERT** [58] is built through a stack of transformer layers, which can easily align text and images by using the self-attention mechanism.
- **CLIP** [59] is a contrastive language-image pretraining model that leverages natural language supervision to learn visual concepts, building upon and simplifying ConVIRT.
- **MOMENTA** [44] is a multimodal neural network that can utilize global and local information from input features to enhance model performance.
- **PVLM** [60] is a few-shot multimodal learning method with prompts for modeling visuo-perceptual language.
- **Prompt Approach**, Ji et al. [61] proposed a prompt-based approach to detect harmful memes and boost performance by converting visual cues into textual features.
- **Harmonic-NAS** [62] is a hardware-aware approach for jointly optimizing unimodal backbone and multimodal fusion networks.
- **PMF-large** [63] is an efficient multimodal fusion method devoted to fusing unimodal pre-trained transformers.
- **DynMM** [64] is a novel method for adaptively fusing multimodal data and generating data-dependent forward paths in the inference process.
- **DC-NAS** [65] is an evolutionary-based MNAS approach that achieves time reduction and performance improvement through a divide-and-conquer network structure.
- **MM-ENAS** [66] is a multimodal multi-scale evolutionary network structure search approach that achieves the unified hierarchical feature representation and the optimal fusion operation selection through a two-stage manner.
- **I3D**, Carreira and Zisserman [67] proposed a new Two-Stream Inflated 3D ConvNet for action recognition.
- **MTUT**, Gupta et al. [68] proposed a simple but effective multi-task learning framework to model gesture progression and frame-level recognition.
- **EDF**, Liang et al. [69] proposed an evolutionary algorithm for searching the optimal combination scheme of different fusion operators to fuse multi-view features.
- **CSG-NAS** [70] is an effective MNAS method based on shrink-and-expansion search space concepts and employs an adaptive strategy with evolutionary algorithms to facilitate knowledge sharing and reuse.

C. Experimental Results

1) Experimental Results on HARM_P and HARM_C Datasets: As shown in Table II, HF-MNAS obtains the optimal results on the harmful meme detection tasks across unimodal and multimodal. Specifically, on the HARM_P dataset (3)-Class classification), HF-MNAS is superior to BERT by 19.93% and 1.87% in terms of the F1-M score and accuracy when trained with text modality. Our model outperforms ResNeXt-101 by 9.65% and 2.43% in the F1-M score and accuracy under image-only training. Compared to multimodal fusion methods, our method outperforms all of them. For instance, our model exceeds the method proposed by [61] in terms of F1-M score and accuracy by 2.1% and 1.83%. Our method outperforms BM-NAS by 5.32% and 3.37% in F1-M score and accuracy. On the HARM_C dataset (3)-Class classification), HF-MANS also achieved consistent performance improvements. For instance, our proposed method is superior to BERT by 2.87% and 3.51%, to ResNeXt-101 by 7.17% and 0.34%, and to BM-NAS by 4.03% and 2.01% in F1-M score and accuracy. For the HARM_C and HARM_P datasets with the 2-class classification mentioned in Table II, our model still achieves a comparable experimental result. From these encouraging results, it is clear that our model is able to better fuse the relevant information between different modalities. Moreover, the ablation study in Section V-D also provides sufficient evidence for the validity of our method.

2) Experimental Results on MM-Imbd Dataset: Table III reports the results of HF-MNAS on the MM-IMDB dataset compared with the SOTA methods. It is clear that HF-MNAS consistently outperforms baselines in most test cases. Specifically, HF-MNAS outperforms Maxout MLP by 4.70% and 14.87% in terms of the F1-W and F1-M scores, respectively, when trained with text modality. The proposed method is

TABLE II

EXPERIMENTAL RESULTS WITH MACRO F1 (F1-M) AND ACCURACY ON HARM_P DATASET AND HARM_C DATASET. † INDICATES EXPERIMENTAL RESULTS OF OWN REPRODUCTION. THE 3-CLASS CLASSIFICATION DATASETS CONTAIN VERY HARMFUL, PARTIALLY HARMFUL, AND UNHARMFUL CATEGORIES. THE 2-CLASS CLASSIFICATION DATASETS CONTAIN HARMFUL AND UNHARMFUL CATEGORIES. THE RED INDICATES THE BEST RESULTS AND THE GREEN INDICATES THE SECOND BEST

	Method	Modalities	HARM_P Dataset				HARM_C Dataset			
			2-Class Classification		3-Class Classification		2-Class Classification		3-Class Classification	
			F1-M (%)	Acc (%)	F1-M (%)	Acc (%)	F1-M (%)	Acc (%)	F1-M (%)	Acc (%)
Unimodal	BERT	T	78.35	80.12	54.08	74.55	66.25	70.17	48.72	68.93
	BERT+Prompt	T	- -	- -	55.09	55.37	- -	- -	53.59	71.88
	HF-MNAS (Our)	T	78.99	81.80	74.01	76.42	70.48	75.33	64.52	72.36
	VGG	I	70.46	70.65	51.89	73.65	61.86	68.12	41.76	66.24
	DenseNet-161	I	73.68	74.05	50.98	71.80	62.54	68.42	42.15	65.21
	ResNet-152	I	72.77	73.14	50.64	71.02	62.97	68.74	43.02	65.29
	ResNeXt-101	I	73.57	73.91	51.45	71.84	63.68	69.79	43.68	66.55
	HF-MNAS (Our)	I	75.30	74.71	61.10	74.27	67.75	73.57	50.85	66.89
Multimodal	MMBT	I+T	80.23	82.54	58.03	78.14	67.12	73.48	50.88	68.08
	ViLBERT	I+T	86.03	87.25	64.70	84.66	78.06	78.53	48.82	75.71
	VisualBERT	I+T	86.07	86.80	63.68	84.02	80.13	81.36	53.85	74.01
	CLIP	I+T	80.25	80.55	56.85	77.00	73.85	74.23	44.25	67.04
	CLIP + Proposals	I+T	83.80	84.16	60.65	81.06	76.90	77.65	45.60	70.52
	CLIP + Attributes	I+T	83.85	84.02	60.23	80.75	77.64	78.10	45.55	71.05
	MOMENTA	I+T	88.26	89.84	66.66	87.14	82.80	83.82	54.74	77.10
	PVLM	I+T	- -	- -	65.72	67.01	- -	- -	38.96	59.74
	BM-NAS†	I+T	87.21	88.73	86.42	87.19	81.81	83.65	74.03	81.75
	Prompt Approach	I+T	- -	- -	89.64	88.73	- -	- -	54.74	77.10
	HF-MNAS (Our)	I+T	89.08	90.34	91.74	90.56	84.16	85.41	78.06	83.76
	Improvement	- -	**1.87 ↑**	**1.61 ↑**	**5.32 ↑**	**3.37 ↑**	**2.35 ↑**	**1.76 ↑**	**4.03 ↑**	**2.01 ↑**

higher than VGG by 0.62% and 0.57% in terms of the F1-W and F1-M scores under image-only training. This result shows that our method could select the appropriate operation for different modalities to boost the performance of tasks. Compared to the NAS-based approaches, our method outperforms MFAS and BM-NAS by 1.67% and 1.60% in F1-W and by 6.99% and 0.98% in F1-M, respectively. Compared with CSG-NAS, DC-NAS, and MM-ENAS methods, HF-NAS still shows comparable performance. The performance improvement can be explained that our hierarchical fusion method can bring better interaction between modalities. Moreover, the use of DCCA and multi-head attention can further mitigate the interference of modality inconsistency.

3) Experimental Results on NTU RGB-D and EgoGesture Datasets: To validate the generalizability of HF-MNAS, we compare the experimental results of NTU RGB-D (involving video and pose modalities) and EgoGesture (containing RGB and depth image modalities) multimodal datasets. As shown in Tables IV and V, our proposed method consistently outperforms the baseline models across most test cases. Specifically, on NTU RGB-D dataset, when both video and pose modalities are used for training, HF-MNAS achieves 0.67% higher than BM-NAS in Acc metrics (1.65% higher than MFAS, and 0.3% higher than DC-NAS, respectively). On the EgoGesture dataset, when both RGB and Depth image are used for training, HF-MNAS achieves 0.35% higher than BM-NAS in Acc metrics (0.15% higher than EDF, and 0.09% higher than DC-NAS, respectively). Then, it is clear that our proposed hierarchical Fusion multimodal neural architecture search method exhibits strong generalizability and can effectively seek competitive fusion architectures for multimodal tasks.

D. Ablation Study

1) Impact of Different Components: To systematically evaluate the contribution of each component in our model, we conducted extensive ablation experiments on the HARM_P (3)-Class classification) and MM-IMDB datasets in both unimodal and multimodal settings. As shown in Tables VI and VII, the base architecture comprises three fundamental modules: UFEM, LFBC, and HFBC. Through incremental integration of core components (DCCA, MHA, and LSTM), we demonstrate that each component contributes positively to model performance. Notably, the complete architecture integrating all components (Base+DCCA+MHA+LSTM), designated as our HF-MNAS framework, achieves optimal performance. Specifically, on the HARM_P dataset, HF-MNAS achieves a 4.14% and 4.73% improvement over Base in F1-M and Acc scores under text-only training. For image-only

TABLE III

EXPERIMENTAL RESULTS WITH WEIGHTED F1 (F1-W) AND MACRO F1 (F1-M) ON MM-IMDB. * DENOTES MNAS METHODS, AND † INDICATES EXPERIMENTAL RESULTS OF OWN REPRODUCTION

	Method	Modalities	F1-W(%)	F1-M(%)
Unimodal	Maxout MLP	T	57.54	45.98
	BERT	T	- -	59.60
	Harmonic-NAS*	T	61.18	- -
	HF-MNAS (Our)	T	62.24 (↑4.70)	60.85 (↑14.87)
	VGG	I	49.21	33.50
	ViT	I	- -	32.20
	Harmonic-NAS*	I	46.12	- -
	HF-MNAS (Our)	I	49.83 (↑0.62)	34.07 (↑0.57)
Multimodal	Two-stream	I+T	60.81	50.49
	GMU	I+T	61.70	54.10
	CentralNet	I+T	62.23	53.44
	MFAS*	I+T	62.50	55.68
	MMBT	I+T	- -	61.40
	BM-NAS*	I+T	62.92	- -
	BM-NAS†*	I+T	62.57	61.69
	MMBV	I+T	- -	61.70
	DynMM	I+T	- -	51.60
	Harmonic-NAS*	I+T	63.61	- -
	CSG-NAS*	I+T	64.12	- -
	DC-NAS*	I+T	63.70	- -
	PMF-large	I+T	- -	61.66
	MM-ENAS*	I+T	64.01	- -
	HF-MNAS (Our)	I+T	64.17 (↑1.60)	62.67 (↑0.98)

TABLE IV

EXPERIMENTAL RESULTS ON NTU RGB-D DATASET

Setting	Method	Modality	Acc(%)
Unimodal	Inflated ResNet-50	V	83.91
	Co-occurence	P	85.24
Multimodal	Two-stream	V+P	88.60
	GMU	V+P	85.80
	MMTM	V+P	88.92
	CentralNet	V+P	89.36
	MFAS*	V+P	89.50
	EDF*	V+P	91.08
	BM-NAS*	V+P	90.48
	DC-NAS*	V+P	90.85
	CSG-NAS*	V+P	91.12
	HF-MNAS (Ours)	V+P	91.15

TABLE V

EXPERIMENTAL RESULTS ON EGOGESTURE DATASET

Setting	Method	Modality	Acc(%)
Unimodal	I3D	R	90.33
	ResNext-101	R	93.75
	I3D	D	89.47
	ResNext-101	D	94.03
Multimodal	I3D	R+D	92.78
	MMTM	R+D	93.51
	MTUT	R+D	93.87
	EDF*	R+D	95.16
	BM-NAS*	R+D	94.96
	DC-NAS*	R+D	95.22
	CSG-NAS*	R+D	95.25
	HF-MNAS (ours)	R+D	95.31

TABLE VI

COMPARISON AMONG DIFFERENT COMPONENTS ON TOP OF THE BASE METHOD (CONSISTING OF UFEM +LFBC +HFBC) IN THE UNIMODAL SETTING ON THE HARM_P DATASET

Unimodal settings	Text		Image	
	F1-M(%)	Acc(%)	F1-M(%)	Acc(%)
Base	69.87	71.69	58.15	70.32
Base +MHA	72.79	74.33	60.32	73.18
Base +DCCA	72.64	73.94	60.03	72.47
Base +DCCA +MHA	73.26	75.11	61.10	74.27
Base +DCCA +MHA +LSTM	74.01	76.42	- -	- -

TABLE VII

COMPARISON AMONG DIFFERENT COMPONENTS ON TOP OF THE BASE METHOD (CONSISTING OF UFEM +LFBC +HFBC) IN THE MULTIMODAL SETTING ON MM-IMDB AND HARM_P DATASETS

Multimodal settings	MM-IMDB	HARM_P	
	F1-W(%)	F1-M(%)	Acc(%)
Base	62.22	87.66	86.93
Base +MHA	62.59	89.62	88.06
Base +DCCA	62.40	89.15	87.22
Base +DCCA +MHA	63.32	90.03	88.61
Base +DCCA +MHA +LSTM	64.17	91.74	90.56

training, HF-MNAS surpasses the base by 2.95% and 3.95% in F1-M and Acc scores. When both image and text modalities are used for training, HF-MNAS achieves 4.08% and 3.63% higher than Base in terms of F1-M and Acc metrics. Additionally, the core components also show tremendous potential on the MM-IMDB dataset. It is a strong demonstration that our employed components can improve the model's overall performance.

2) Impact of Unimodal Features Selection: In order to validate the effectiveness of fusing low-level features using LFBC and high-level features using HFBC, we perform a set of ablation experiments at the feature inputs of LFBC and HFBC on the HARM_P (3)-Class classification) dataset. As shown in Fig. 10, the abscissa indicates the input features

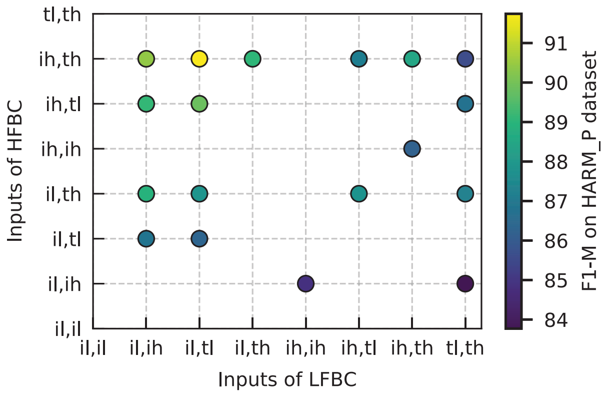


Fig. 10. The impact of unimodal feature selection.

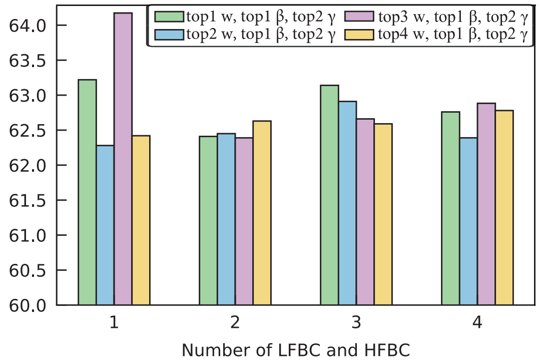


Fig. 11. Impact of hyperparameters (β, γ, and w and the stacking number of LFBC and HFBC on model performance on the MM-IMDB dataset.

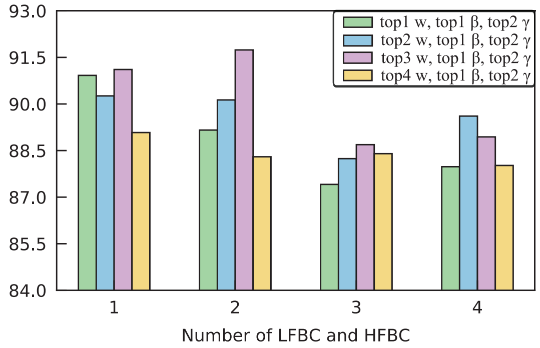


Fig. 12. Impact of hyperparameters (β, γ, and w and the stacking number of LFBC and HFBC on model performance on the HARM _P dataset.

of LFBC, and the verticle represents the input features of HFBC. The different color dots indicate the F1-M scores. For instance, the yellow dot indicates that the F1-M score is 91.74% on the HARM _P dataset when the inputs of LFBC are low-level text (*tl*) and image features (*il*) and the inputs of HFBC are high-level text (*th*) and image features (*ih*), where *i* and *t* represent text and image, *l* and *h* indicate low-level and high-level features, respectively. From Fig. 10, we can see that an unexpected result can be achieved by using two independent cells to fuse the low- and high-level features of different modalities.

3) Hyper-Parameters Analysis: We then investigate the impact of hyperparameters (β, γ, and w) on the performance of the multimodal task. To systematically evaluate their influence, we fix the number of intermediate nodes in both LFBC and HFBC cells to 4 during the search phase. Subsequently, we employ a continuous relaxation strategy to alternately optimize the architectural parameters (β and γ) and network weight parameters (w). In the inference phase, we select the operation with the largest architectural weight (i.e., top1 β) as the fusion operation of the intermediate node, and the features with the top 2 architectural weights (i.e., top2 γ) as the inputs of the intermediate node. Next, according to the network weights w, we select the top K most important intermediate node outputs (e.g., top3 w) cascade as the output features of the LFBC or HFBC cells. Finally, we adjust the stacking number of LFBC and HFBC cells to analyze the impact of hyperparameters on the model performance.

TABLE VIII
THE EFFECT OF CANDIDATE OPERATIONS OF INTERMEDIATE NODES

Method	MM-IMDB	HARM_P	
	F1-W(%)	F1-M(%)	Acc(%)
Sum	62.52	**89.48**	84.65
ScaleDotAttention	62.49	86.05	83.24
Bidirectional_Attention	63.24	87.78	86.64
Squeeze_Excitation	62.67	88.11	84.94
Multi-head Attention	63.31	89.45	85.22
LinearGLU	62.52	88.28	83.24
ConcatFC	62.29	88.87	**87.22**
Multiply	62.23	88.64	83.78
Mamba_fuison	**63.37**	89.12	86.98

TABLE IX
THE COMPARISON OF SEARCH COST (GPU-HOURS)

Method	MM-IMDB		
	Dev (F1-M%)	Test (F1-M %)	Search cost
MFAS	--	62.50 (+ 1.67)	9.24 (↑ 12.8x)
BM-NAS	53.52 (+ 1.07)	62.57 (+ 1.60)	0.89 (↑ 1.24x)
HF-MNAS	**54.59**	**64.17**	**0.72**

As shown in Figs. 11 and 12, on the MM-IMDB dataset, the optimal performance is achieved when the stacking number of LFBCs and HFBCs is 1, and the architectural and network parameters are top1 β, top2 γ, and top3 w; on the HARM _P (3)-class classification) dataset, the highest performance is gained when the number of stacks of LFBCs and HFBCs is 2, and the architectural and network parameters are top1 β, top2 γ, and top3 w. Therefore, from Figs. 11 and 12, we can find that the selection of the architectural parameters (β and γ), network weight parameters (w) and the stacked number of cells indeed influence the model performance to a different extent. This is due to the fact that the intermediate nodes and the stacked number of cells can capture different information during feature fusion thus leading to different results.

4) Search Cost: Table IX provides a comprehensive comparison of the search cost between the NAS-based approaches on the MM-IMDB dataset, where HF-MNAS only utilizes one LFBC and HFBC and the number of intermediate nodes is 4. As shown in table IX, compared with MFAS, our method achieves a search speed at least 12x faster and also improves the F1-M score by 1.67%. Compared to BM-NAS, HF-MNAS achieves at least 1.24x search speed-up and 1.60% enhancement in F1-M score.

5) Impact of Intermediate Node Operations: To explore the impact of candidate fusion operations of intermediate nodes on the model performance, in the ablation experimental settings, we fix the stacking number of LFBCs and HFBCs, as well as the number of intermediate nodes, to be 1. Subsequently, we conduct extensive experiments on the MM-IMDB and HARM _P (3)-class classification) datasets and evaluate the specific contribution of each candidate fusion operation to overall model performance. As shown in Table VIII, on the MM-IMDB dataset, the Mamba _fusion and Multi-head Attention operations demonstrate superior efficacy; on the

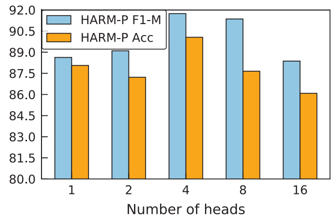


Fig. 13. Impact of head numbers on HARM_P dataset.

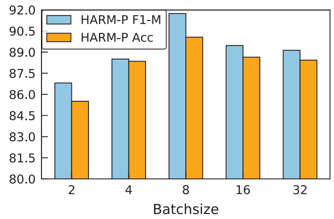


Fig. 14. Impact of Batch size on HARM_P dataset.

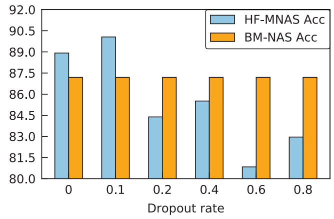


Fig. 15. Impact of dropout rate on accuracy.

HARM_P dataset, the Sum, Multi-head Attention, and ConcatFC operations show competitive performance. Notably, our carefully designed search space offers significant extensibility, allowing flexible integration of advanced fusion methods (e.g., Mamba_fusion) via feature dimension adjustment to further optimize the fusion architecture.

Furthermore, the observed performance disparity between HARM_P and MM-IMDB datasets primarily arises from the fact that HARM_P represents a simpler 3-class classification task, while MM-IMDB involves more challenging multi-label classification across 26 fine-grained categories with frequent label co-occurrences. This difference also highlights the sensitivity of different tasks to the fusion operation selection. Thus, it is possible to integrate more potential fusion operations in the designed search space to improve the classification accuracy for different tasks in the future.

6) Impact of Head Numbers of Multi-Head Attention: To evaluate the impact of the head numbers of multi-head attention in multimodal fusion, we conduct experiments using different numbers of heads on HARM_P (3)-class classification). As shown in Fig. 13, the blue and yellow denote the F1-M and accuracy, respectively. From the figure, we observe that different head numbers generate different impact degrees over the model performance, e.g., when the head number is 4, the model performs best.

7) Impact of Batch Size and Dropout Rate: We explore the effect of batch size and dropout rate on model performance on HARM_P (3)-class classification). As shown in Fig. 14, the blue indicates the F1-M and the yellow denotes the accuracy. As shown, as the batch size increases, the model performance tends to increase and then decrease slowly, the best performance is achieved when the batch size is 8. Fig. 15 and 16 show the simulation results of HF-MNAS and BM-NAS in terms of the impact of dropout rate, where the blue indicates the accuracy and F1-M achieved by our method for different dropout rate settings. From the figures, we find that the model achieves a great performance when dropout is set to 0.2.

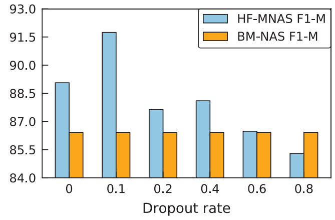


Fig. 16. Impact of dropout rate on F1-M.

VI. Conclusion

The goal of this paper is to design an efficient multimodal feature fusion approach for multimodal learning tasks. This goal is realized by the proposed hierarchical fusion architecture search method with an inconsistency mitigation strategy (called HF-MNAS). Different from existing approaches, our HF-MNAS performs fine-grained exploitation in multi-level fusion architectures, and tackles the issue of inconsistency in modalities and labels. The proposed HF-MNAS method is examined on a set of multimodal datasets. The experimental results validate the competitiveness of HF-MNAS in dealing with various multimodal learning tasks, the efficiency of the proposed search space and optimization strategies, and the effectiveness of the designed components in HF-MNAS.

However, the proposed HF-MNAS approach can still be improved in two aspects. First, the search space size of multimodal feature fusion can be expanded. In particular, the cell-based search space in HF-MNAS only contains seven optional fusion operations and is dominated by attention mechanisms, which limits the potential of seeking more prospective architectures. However, we should notice that an enlarged search space often leads to an increase in computational cost. Second, the feature extraction ability of the unimodal backbone model can be further enhanced. In fact, HF-MNAS employs the common unimodal backbone model and yet achieves superior performance. It is still necessary to integrate advanced unimodal feature extraction methods into our method for further performance improvement. In the future, we will pay more attention to the holistic study of the enlarged yet efficient search space as well as the enhanced backbone model for compound benefits.

References

[1] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "CentralNet: A multilayer approach for multimodal fusion," in *Proc. ECCV*, 2019, pp. 575–589.

[2] T. Jin, S. Huang, M. Chen, Y. Li, and Z. Zhang, "SBAT: Video captioning with sparse boundary-aware transformer," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 630–636.

[3] J.-M. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "MFAS: Multimodal fusion architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6959–6968.

[4] Y. Yin, S. Huang, and X. Zhang, "BM-NAS: Bilevel multimodal neural architecture search," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 8, pp. 8901–8909.

[5] J. Zhang and W. Li, "Multi-modal and multi-scale temporal fusion architecture search for audio-visual video parsing," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 3328–3336.

[6] M.-I. Georgescu et al., "Multimodal multi-head convolutional attention with various kernel sizes for medical image super-resolution," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2195–2205.

[7] X. Hu, J. Jiang, C. Wang, X. Liu, and J. Ma, "Incrementally adapting pretrained model using network prior for multi-focus image fusion," *IEEE Trans. Image Process.*, vol. 33, pp. 3950–3963, 2024.

[8] H. Liu, Z. Ni, D. Nie, D. Shen, J. Wang, and Z. Tang, "Multimodal brain tumor segmentation boosted by monomodal normal brain images," *IEEE Trans. Image Process.*, vol. 33, pp. 1199–1210, 2024.

[9] L.-A. Zeng and W.-S. Zheng, "Multimodal action quality assessment," *IEEE Trans. Image Process.*, vol. 33, pp. 1600–1613, 2024.

[10] J. Arévalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated multimodal units for information fusion," in *Proc. ICLR*, 2017.

[11] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," 2017, *ArXiv:1802.00924*.

[12] D. H. Kim, M. K. Lee, D. Y. Choi, and B. C. Song, "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 529–535.

[13] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. NeurIPS*, 2019, pp. 13–23.

[14] H. R. Vaezi Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for CNN fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13286–13296.

[15] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," in *Proc. ViGIL-NeurIPS*, 2019.

[16] B. Baker, O. Gupta, N. Naik, and R. Raskar, "Designing neural network architectures using reinforcement learning," in *Proc. ICLR*, 2016.

[17] L. Ma, S. Cheng, and Y. Shi, "Enhancing learning efficiency of brain storm optimization via orthogonal learning design," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 11, pp. 6723–6742, Nov. 2021.

[18] N. Nayman, A. Noy, T. Ridnik, I. Friedman, R. Jin, and L. Zelnik, "XNAS: Neural architecture search with expert advice," in *Proc. Adv. NeurIPS*, vol. 32, 2019, pp. 1975–1985.

[19] H. Yu et al., "Cyclic differentiable architecture search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 211–228, Jan. 2023.

[20] L. Ma et al., "Defying multi-model forgetting in one-shot neural architecture search using orthogonal gradient learning," *IEEE Trans. Comput.*, vol. 74, no. 5, pp. 1678–1689, May 2025.

[21] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proc. 30th Int. Conf. Mach. Learn.*, Feb. 2013, pp. 115–123.

[22] H. Cai, L. Zhu, and S. Han, "ProxylessNAS: Direct neural architecture search on target task and hardware," in *Proc. ICLR*, 2018.

[23] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.

[24] H. Liu, K. Simonyan, O. Vinyals, C. Fernando, and K. Kavukcuoglu, "Hierarchical representations for efficient architecture search," 2017, *arXiv:1711.00436*.

[25] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 4780–4789.

[26] L. Ma et al., "A novel fuzzy neural network architecture search framework for defect recognition with uncertainties," *IEEE Trans. Fuzzy Syst.*, vol. 32, no. 5, pp. 3274–3285, May 2024.

[27] G. Yuan, B. Wang, B. Xue, and M. Zhang, "Particle swarm optimization for efficiently evolving deep convolutional neural networks using an autoencoder-based encoding strategy," *IEEE Trans. Evol. Comput.*, vol. 28, no. 5, pp. 1190–1204, Oct. 2024.

[28] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in *Proc. ICLR*, 2018.

[29] Y. Peng, L. Bi, M. Fulham, D. Feng, and J. Kim, "Multi-modality information fusion for radiomics-based neural architecture search," in *Proc. MICCAI*, 2020, pp. 763–771.

[30] Z. Yu, Y. Cui, J. Yu, M. Wang, D. Tao, and Q. Tian, "Deep multimodal neural architecture search," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3743–3752.

[31] P. Fu, X. Liang, Y. Qian, Q. Guo, Z. Wei, and W. Li, "CoMO-NAS: Core-structures-guided multi-objective neural architecture search for multi-modal classification," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 9126–9135.

[32] Z. Xu, D. R. So, and A. M. Dai, "MUFASA: Multimodal fusion architecture search for electronic health records," in *Proc. AAAI*, vol. 35, 2021, pp. 10532–10540.

[33] T. Li, X. Yang, Y. Ke, B. Wang, Y. Liu, and J. Xu, "Alleviating the inconsistency of multimodal data in cross-modal retrieval," in *Proc. IEEE 40th Int. Conf. Data Eng. (ICDE)*, May 2024, pp. 4643–4656.

[34] D. Cheng, H. Tai, N. Wang, C. Fang, and X. Gao, "Neighbor consistency and global–local interaction: A novel pseudo-label refinement approach for unsupervised person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 9070–9084, 2024.

[35] J. Chen, W. Huang, J. Zhang, K. Debattista, and J. Han, "Addressing inconsistent labeling with cross image matching for scribble-based medical image segmentation," *IEEE Trans. Image Process.*, vol. 34, pp. 842–853, 2025.

[36] S. Zhao, Y. Liu, Q. Jiao, Q. Zhang, and J. Han, "Mitigating modality discrepancies for RGB-T semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9380–9394, Jul. 2024.

[37] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," in *Classic Works of the Dempster–Shafer Theory of Belief Function*, vol. 38, Berlin, Germany: Springer, 2008, pp. 325–339.

[38] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2551–2566, Feb. 2023.

[39] X. Liang, P. Fu, Y. Qian, Q. Guo, and G. Liu, "Trusted multi-view classification via evolutionary multi-view fusion," in *Proc. ICLR*, 2025.

[40] D. Cheng et al., "Continual all-in-one adverse weather removal with knowledge replay on a unified network structure," *IEEE Trans. Multimedia*, vol. 26, pp. 8184–8196, 2024.

[41] X. Xie, Y. Cui, T. Tan, X. Zheng, and Z. Yu, "FusionMamba: Dynamic feature enhancement for multimodal image fusion with mamba," *Vis. Intell.*, vol. 2, no. 1, p. 37, Dec. 2024.

[42] A. Gudmalwar, B. Basel, A. Dutta, and C. V. Rama Rao, "The magnitude and phase based speech representation learning using autoencoder for classifying speech emotions using deep canonical correlation analysis," in *Proc. Interspeech*, Sep. 2022, pp. 1163–1167.

[43] D. Cheng, L. He, N. Wang, D. Zhang, and X. Gao, "Semantic-aligned learning with collaborative refinement for unsupervised VI-ReID," *Int. J. Comput. Vis.*, vol. 2025, pp. 1–23, May 2025.

[44] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, and T. Chakraborty, "MOMENTA: A multimodal framework for detecting harmful memes and their targets," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2021, pp. 4439–4455.

[45] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[46] Y. Zhang, C. Cao, J. Cheng, and H. Lu, "EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1038–1050, May 2018.

[47] I. Goodfellow et al., "Maxout networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1319–1327.

[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[49] F. Petroni et al., "Language models as knowledge bases?," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 2463–2473.

[50] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[51] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 4700–4708.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[53] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 1492–1500.

[54] F. Baradel, C. Wolf, J. Mille, and G. W. Taylor, "Glimpse clouds: Human activity recognition from unstructured feature points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 469–478.

[55] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 786–792.

[56] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. NeurIPS*, 2014, pp. 568–576.

[57] I. Monter-Aldana, A. P. Lopez Monroy, and F. Sanchez-Vega, "Dynamic regularization in UDA for transformers in multimodal classification," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 8700–8711.

[58] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*.

[59] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.

[60] Y. Yu and D. Zhang, "Few-shot multi-modal sentiment analysis with prompt-based vision-aware language modeling," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

[61] J. Ji, W. Ren, and U. Naseem, "Identifying creative harmful memes via prompt based approach," in *Proc. ACM Web Conf.*, Apr. 2023, pp. 3868–3872.

[62] M. I. E. Ghebriout, H. Bouzidi, S. Niar, and H. Ouarnoughi, "Harmonic-NAS: Hardware-aware multimodal neural architecture search on resource-constrained devices," in *Proc. ACML*, 2023, pp. 374–389.

[63] Y. Li, R. Quan, L. Zhu, and Y. Yang, "Efficient multimodal fusion via interactive prompting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2604–2613.

[64] Z. Xue and R. Marculescu, "Dynamic multimodal fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2575–2584.

[65] X. Liang, P. Fu, Q. Guo, K. Zheng, and Y. Qian, "DC-NAS: Divide-and-conquer neural architecture search for multi-modal classification," in *Proc. AAAI*, 2024, vol. 38, no. 12, pp. 13754–13762.

[66] P. Fu et al., "Multi-scale features are effective for multi-modal classification: An architecture search viewpoint," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 2, pp. 1070–1083, Feb. 2025.

[67] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[68] V. Gupta, S. K. Dwivedi, R. Dabral, and A. Jain, "Progression modelling for online and early gesture detection," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 289–297.

[69] X. Liang, Q. Guo, Y. Qian, W. Ding, and Q. Zhang, "Evolutionary deep fusion method and its application in chemical structure recognition," *IEEE Trans. Evol. Comput.*, vol. 25, no. 5, pp. 883–893, Oct. 2021.

[70] P. Fu, X. Liang, T. Luo, Q. Guo, Y. Zhang, and Y. Qian, "Core-structures-guided multi-modal classification neural architecture search," in *Proc. 33rd Int. Joint Conf. Artif. Intell.*, Aug. 2024, pp. 3980–3988.

Kaifang Long received the B.Sc. degree from Taishan University, Taian, in 2020, and the M.Sc. degree from Shandong Normal University, Jinan, China, in 2023. He is currently pursuing the Ph.D. degree with the College of Software, Northeastern University, Shenyang, China. His research interests include multimodal learning and anomaly detection.

Guoyang Xie (Member, IEEE) received the B.Sc. degree from the University of Electronic Science and Technology of China in 2013, the MPhil. degree from The Hong Kong University of Science and Technology in 2015, and the Ph.D. degree from the University of Surrey, Guildford, U.K., in 2023. He is currently an Algorithm Manager at CATL, Ningde, China. His research interests include AI for manufacturing and medical imaging, industrial image anomaly detection, robot learning, and data synthesis.

Lianbo Ma (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from Northeastern University, Shenyang, China, in 2004 and 2007, respectively, and the Ph.D. degree in mechanical and electronic engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2015. He is currently a Professor with Northeastern University. He has published over 120 journal articles, books, and refereed conference papers. His current research interests include computational intelligence and machine learning.

Qing Li (Senior Member, IEEE) received the B.Sc. degree in computer science and technology from Dalian University of Technology, Dalian, China, in 2008, and the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2013. He is currently an Associate Researcher with the Peng Cheng Laboratory, Shenzhen, China. His research interests include network function virtualization, network caching/computing, intelligent self-running networks, and edge computing.

Min Huang (Member, IEEE) received the B.Sc. degree in automatic instruments, the M.Sc. degree in systems engineering, and the Ph.D. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 1990, 1993, and 1999, respectively.

She is currently a Professor with Northeastern University. She has published over 100 journal articles, books, and refereed conference papers. Her research interests include modeling and optimization for logistics and supply chain.

Jianhui Lv received the B.S. degree in mathematics and applied mathematics from Jilin Institute of Chemical Technology, Jilin, China, in 2012, and the M.S. and Ph.D. degrees in computer science from Northeastern University, Shenyang, China, in 2014 and 2017, respectively.

He is currently a Professor at Jinzhou Medical University, China. His research interests include computer networks, artificial intelligence, ICN, the IoT, bio-inspired networking, evolutionary computation, cloud/edge computing, smart city, and healthcare.

Zhichao Lu (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Michigan State University, USA, in 2020. He is currently an Assistant Professor at the Department of Computer Science, City University of Hong Kong, Hong Kong, China. His current research interests include the intersections of evolutionary computation, learning, and optimization, notably on developing efficient, reliable, and automated machine learning algorithms and systems.