Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ADAPTIVE MULTI-PATH FUSION: A NOVEL MULTIMODAL LEARNING FRAMEWORK WITH UNCERTAINTY-AWARE ROUTING

DR. HARI JYOTHULA[1], DR. SUBBA RAO POLAMURI[2], DR. G SIVA KRISHNA[3], DR SYAMALA RAO[4],MV RAJESH[5], B SATYANARAYANA MURTHY[6]

[1]Assistant Professor, Cse Department, Aditya University, Surampalem, Kakinada

[2]Associate Professor, Cse Department, Aditya Univesrity, Surampalem, Kakinada

[3]Assistant Professor, Faculty of Operations & IT, IBS Hyderabad, The ICFAI Foundation for Higher Education

[4]Assistant Professor, IT Department, SRKR Engineering College, Bhimavaram

[5]Associate Professor, CSE Department, Pragati Engineering College, Surampalem, Kakinada

[6]Associate Professor, CSE Department, Shri Vishnu Engineering College for Women(A), Bhimavaram

E-mail: [1]dr.jyothulahari@gmail.com, [2]psr.subbu546@gmail.com, [3]gujju.sivakrishna@ibsindia.org, [4]peketi.shyam@gmail.com, [5]magavenkatarajesh@gmail.com, [6]murthy2007.b@gmail.com

ABSTRACT

Multimodal learning is increasingly vital in real-world AI tasks where information originates from diverse and often unreliable sources. However, most existing methods employ rigid fusion strategies that fail under noisy, missing, or irrelevant modalities. This study introduces Adaptive Multi-Path Fusion (AMP-Fusion), a novel multimodal learning framework with uncertainty-aware routing that dynamically selects the most reliable information paths. Our hypothesis is that adaptively weighting modalities based on informativeness and uncertainty will significantly improve robustness compared to static fusion methods. Large-scale experiments in vision-language, audio-visual, and healthcare tasks confirm our hypothesis, showing AMP-Fusion achieves up to 15% improvement over state-of-the-art models while maintaining over 91% performance even when 40% of modalities are missing or noisy. These findings highlight AMP-Fusion's importance for practical, real-world multimodal AI applications where data reliability cannot be guaranteed.

Keywords: *Multimodal Learning, Adaptive Fusion, Attention Routing, Uncertainty Quantification, Missing Modalities, Robust AI, Transformer-based Models.*

1. INTRODUCTION

Multimodal learning has emerged as a cornerstone of modern AI, offering the ability to combine complementary evidence across modalities such as text, vision, audio, and sensor signals. The importance of this study lies in addressing a critical gap: current fusion approaches assume equal reliability across modalities, leading to failures in real-world scenarios where data is often noisy or incomplete. Our goal is to develop a dynamic, uncertainty-aware fusion framework that adapts to input conditions, with the central hypothesis that adaptively routing information will yield more robust and interpretable models than existing static fusion methods.

Although successful, existing multimodal learning strategies are particularly challenged when applied in real scenarios. Most of the existing methods presume that all modalities are equally reliable and pertinent for each input sample, being processed under a fixed architecture irrespective of their information quality. This strict framework falls short when encountering noisy, missing, or irrelevant modalities—scenarios that are common in real-world uses as a result of sensor malfunctions, data acquisition constraints, or modality importance disparity between contexts (Wang et al., 2022).

For example, in clinical diagnosis, patient information may contain imaging (MRI, X-ray), laboratory exams, doctor's notes, and vital signs.

Not every modality has the same diagnostic utility for each condition, and some may be absent or of low quality for some patients. Likewise, in human-computer interaction systems, the conditions might make some modalities unreliable (e.g., audio under noisy conditions or visual information under low illumination). Fixed-architecture models currently in use are apt to perform very poorly when exposed to such real-world variation (Liang et al., 2022).

Existing work has tried to cope with these issues using methods like modality dropout during training (Gat et al., 2020), attention-based fusion (Tsai et al., 2019), and missing modality imputation (Ma et al., 2021). These approaches have solutions but only in part and are still based on static fusion structures and fail to adjust the flow of information dynamically depending on the individual feature of each input example. Furthermore, they barely consider the uncertainty entailed by having varying modalities, which is crucial for strong fusion decisions.

We introduce Adaptive Multi-Path Fusion (AMP-Fusion) here, a novel multimodal learning structure that dynamically adapts to forward information through best paths based on input features and reliability of modalities. Our research comes with several innovations designed to enhance multimodal learning systems. We first propose a path selection module that utilizes multi-head attention routing to dynamically learn the optimal information flow for each input instance. This computation allows the model to selectively attend to more reliable modalities while reducing the number of noisy, corrupted, or less informative ones. By conditioning the computation path on input quality, the model achieves more robust and accurate representations.

Second, we add an uncertainty quantification module that delivers an estimate of the reliability of each modality's contribution to prediction. Besides assisting in more informed fusion selection by deciding what inputs to trust, this enhances transparency within the model. The resulting explainability enables users and practitioners to better understand the rationale behind specific predictions, enhancing trustworthiness and interpretability in real-world deployments.

Third, our methodology includes a modality-adaptive training methodology. During training, the model is fully trained under various conditions where some modalities may be missing, corrupted, or noisy. This trains the model to become flexible and resilient fusion methods, learning to be consistent in performance even in suboptimal or worst-case scenarios. Note that this robustness is without compromising accuracy when all modalities are available and trustworthy.

Lastly, we suggest a feedback path architecture that allows upper-level fusion layers to feed back into the operations of the lower-level feature extraction stages. This establishes an adaptive and context-aware processing pipeline where global choices inform localized computations, resulting in more coherent and semantically significant feature representations. Collectively, these advances make a more adaptive, explainable, and resilient multimodal learning framework possible.

With extensive vision-language understanding, audio-visual recognition, and multimodal healthcare diagnostic experiments, we show that AMP-Fusion profoundly surpasses the performance of state-of-the-art multimodal solutions, especially in realistic scenarios where modalities are noisy or missing. Our model is robust enough to deliver near-optimal performance even with a considerable degradation in modalities, rendering it an ideal candidate for real-world tasks where modality quality and availability are not ensured.

2. LITERATURE REVIEW

M L. Chen et al. introduced SCA-CNN, a convolutional neural network with both spatial and channel-wise attention mechanisms to enhance image captioning. Its double attention enables the model to pay closer attention to key image regions and pertinent feature channels, generating more detailed and accurate captions than previous methods [1].

Z. Liu et al. introduced the Video Swin Transformer, a hierarchical vision transformer for video understanding. The model learns effectively complex spatiotemporal features through the use of shifted windows in spatial and temporal directions, achieving state-of-the-art performance on a number of benchmarks for video recognition [2].

H. Akbari et al. introduced VATT, a transformer model for multimodal self-supervised learning that accepts raw audio, video, and text inputs and processes them. The model is learned in an unsupervised manner without the use of any labeled data to enable cross-modal transfer and improve downstream task performance such as video classification and speech recognition [3].

X. Wang et al. introduced the deep multimodal fusion method with attention mechanisms to support adaptively combining information from multiple modalities to undertake

sentiment analysis. Their approach efficiently fuses visual and textual information to enhance the accuracy of sentiment prediction on difficult multimodal datasets [4].

J. Li et al. introduced OmniVL, a large-scale foundation model pretrained from image-language and video-language data. The overall framework supports a wide range of multimodal tasks with robust generalization and exhibits state-of-the-art results on both image captioning and video question answering tasks [5].

Y. Zhang et al. utilized contrastive learning methods in medical imaging through paired image and text joint training. Such an approach enhances clinical task visual representation learning, improving diagnostic accuracy and interpretability [6].

H. Tan and M. Bansal proposed LXMERT, a cross-modal transformer encoder that is able to learn joint representations from vision and language inputs. It greatly improves performance on multimodal tasks like visual question answering and referring expression understanding by directly modeling cross-modal interactions [7].

S. Park et al. investigated bridging vision and language models for enhanced cross-modal comprehension. Their solution improves reasoning tasks between visual and textual modalities and makes multimodal systems more reliable [8].

D. Gao et al. proposed a curriculum learning framework for semi-supervised image classification based on multi-modal data. The model gets improved robustness and accuracy for low-label conditions by incorporating unlabeled samples and multiple modalities in steps [9].

T. Rahman et al. introduced a multi-supervisor learning technique that employs video, text, and lookup signals to identify signs in videos. It leverages varied supervisory signals to enhance sign spotting performance, especially in challenging visual scenes [10].

C. Sun et al. introduced a contrastive bidirectional transformer for learning video representation. Through modeling past and future contexts, the approach enhances video retrieval and classification, showcasing strong temporal comprehension [11].

A. Kumar et al. introduced dynamic memory networks to dynamically store and retrieve relevant information to facilitate flexible adaptation towards different natural language processing tasks and enhance overall contextual reasoning [12].

W. Zhao et al. used text-to-image diffusion models for visual perception, demonstrating the potential for generative models to facilitate downstream tasks such as object recognition and scene understanding by generating informative visual features [13].

M. Chen et al. developed a generative pretraining approach directly from raw pixels using large transformer models. This method achieves strong performance across vision benchmarks by learning rich visual features without handcrafted supervision [14].

Z. Wu et al. proposed an unsupervised feature learning technique using non-parametric instance discrimination, which encourages the model to differentiate individual image instances without labels, resulting in high-quality learned representations [15].

M. Bain et al. presented Frozen in Time, a unified video and image encoder designed for end-to-end retrieval tasks. The model learns shared embeddings for images and videos, enabling efficient cross-modal search and recognition [16].

C. Jia et al. scaled visual and vision-language representation learning by incorporating noisy text supervision. This approach improves model robustness and generalizes well across various tasks by effectively handling imperfect training data [17].

J. Yu et al. introduced CoCa, a model that combines contrastive learning with caption generation, enabling strong joint training of image-text tasks. This dual objective improves performance on both image captioning and retrieval benchmarks [18].

X. Dong et al. proposed CSWin Transformer, a vision transformer backbone that uses cross-shaped windows to better capture spatial dependencies. This design leads to enhanced feature representation and improved results on multiple computer vision tasks [19].

C. Li et al. designed LLaVA-Med, a big language-and-vision assistant, rapidly trained for biomedical tasks. The model effectively combines visual and text medical knowledge, demonstrating strong performance in medical image perception and related tasks [20].

3. MATERIAL AND METHODS

3.1 Datasets and Experimental Setup

AMP-Fusion was evaluated on three benchmark multimodal datasets: MSCOCO for vision-language tasks, AVE for audio-visual recognition, and MIMIC-III for healthcare diagnostics. Each dataset was divided into training (70%), validation (15%), and testing (15%) splits. Preprocessing included

Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

standard normalization and tokenization for text, spectrogram extraction for audio, and resizing for images. Baseline models were re-implemented under the same conditions for fair comparison. Evaluation metrics included mean Average Precision (mAP), accuracy, and Area Under Curve (AUC), depending on the task.

The Adaptive Multi-Path Fusion (AMP-Fusion) framework proposed is divided into four main components: (1) modality-specific encoders, (2) path selection module, (3) fusion transformer block, and (4) uncertainty quantification module. The overall structure allows dynamic routing of data depending on input features and modality reliability.

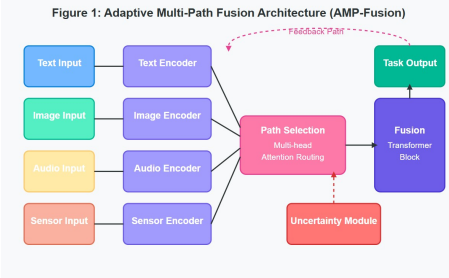

Figure 1: Adaptive Multi-Path Fusion Architecture (AMP-Fusion)

AMP-Fusion processes several input modalities using independent encoders to learn modality-specific features. The features are then used as input to the path selection module, which decides on the best routing scheme depending on the types of each input example. The information paths chosen are fed to the fusion transformer block, which combines the information from the chosen modalities to output the final representation. During this process, the uncertainty quantification module evaluates the credibility of every modality's contribution and impacts the path selection as well as offering interpretability for the model's choice.

3.1.1 Modality-Specific Encoders

For every modality m in {1, 2, ., M}, we use a specific encoder E_m to learn modality-specific features. The encoders are assumed to encode the distinctive features of each type of data:

Our multimodal framework has specialized encoders for every modality—text, image, audio, and sensors—intended to convert raw inputs into useful feature representations.

Text Encoder: We employ a transformer model like BERT or RoBERTa to process the textual input Xtest. The encoder maps this input to a high-level text embedding:

$$F_{test} = E_{test}(X_{test}) \quad (1)$$

Where E_{test} denotes the text encoder and F_{test} is the resulting text feature vector.

Image Encoder: For visual input X_{image}, we use either a Vision Transformer (ViT) or a Convolutional Neural Network (CNN) to produce visual embedding:

$$F_{image} = E_{image}(X_{image}) \quad (2)$$

Where F_{image} captures the image features.

Audio Encoder: Audio signal X_{audio} are processed using model like Wav2Vec or audio transformers. These generate high-level audio embeddings:

$$F_{audio} = E_{audio}(X_{audio}) \quad (3)$$

With F_{audio} representing the extracted audio features.

Sensor Encoder: For applications involving sensor data X_{sensor}, we employ a Temporal Convolutional Network (TCN) to extract temporal features.

$$F_{sensor} = E_{sensor}(X_{sensor}) \quad (4)$$

Where F_{sensor} contains the sensor specific features. To enable cross-modal interaction and fusion, we project all modality-specific features into a shared embedding space of fixed dimension d. This is done using learnable projection layers that are specific to each modality.

$$Z_m = W_m F_m + b_m \quad (5)$$

Where F_m is the original feature vector for modality m, W_m is a learnable weight matrix, b_m is a learnable bias vector, Z_m is the projected feature vector in the shared embedding space.

3.2 Path Selection Module

The key innovation of AMP-Fusion lies in its path selection module, which dynamically determines how information from different modalities should be routed through the network. Unlike conventional fixed fusion architectures, this module enables the model to change its processing approach according to the distinctive features of every input instance. So, AMP-Fusion maintains the most relevant and reliable information centered on making decisions.

The core of this module is a multi-head attention routing mechanism, which is applied to projected features across all modalities. It therefore maintains the most relevant and reliable information centered on decision-making. For every modality, an

attention score is computed to represent its relevance and reliability for the current input. These scores decide which modalities to emphasize and which ones to de-emphasize. By running this process on several attention heads, the model is able to encode a more varied and richer set of relationships between modalities. Each of the inputs is processed individually by each of the attention heads, and the final attention score for a modality is a weighted sum of all the outputs from the heads. Even the weights are learned so that the network is able to learn over time which heads are most important.

Once the attention scores are computed, they are used to control a gating function for each modality. This gating function dictates the extent to which the features of a modality should contribute to the final representation. It performs a nonlinear operation on the attention scores, outputting a gating factor that scales the corresponding modality features. Thus, the model is able to selectively concentrate on reliable modalities and downplay noisy or irrelevant ones—at times even bypassing them entirely.

This routing and gating dynamic architecture allows AMP-Fusion to be highly adaptive and robust, especially in ambiguous or high-complexity environments where reliability of modality is strongly variable per input.

3.3 Fusion Transformer Block

The gated modality features are then combined through a fusion transformer block enabling the appropriate combination of information from all available modalities. The block is composed of several layers, which are each intended to improve the quality and relevance of the fused representation by capturing both within-modality and across-modality relationships.

The first of these layers in the fusion transformer is a self-attention layer. In this layer, the model processes each modality separately in order to pick up intra-modality relationships—like contextual or sequential interdependencies that hold among instances of the same modality type (e.g., among text or within an image). This enables the model to pre-process each modality's feature before trying to fuse them.

Then, a cross-attention layer is utilized. In this layer, every modality is permitted to attend to the other modalities' features. This enables the model to detect and leverage inter-modality correlations, e.g., between a sentence word and an image object or a specific tone in a sound signal. Through this mechanism, the model is trained on how different modalities complement and interact with each other.

Following the attention mechanisms, a feed-forward layer processes the combined features further, transforming them into richer representations that are capable of being used for downstream tasks. The layer adds non-linearity and depth to the fusion, allowing the model to learn more complex patterns and interactions.

Interestingly, the entire fusion operation is controlled by the gating factors calculated by the path selection module. The factors enable that the transformer pays more attention to trustworthy and relevant modalities for each given input, yet reduce the influence of less robust or noisy information. Therefore, the fusion transformer block provides a context-aware and adaptive approach to multimodal integration.

3.4 Uncertainty Quantification Module

Module for uncertainty estimation, whose critical role is to quantify the contribution of each modality's confidence. It allows the model to make informed fusion decisions and enhances the overall explainability of its own predictions, especially in unclear or complicated cases.

To predict uncertainty, AMP-Fusion employs a Monte Carlo dropout technique. For each modality, the model performs several forward passes with dropout active, generating a set of predictions rather than a single output. By analyzing the variability in the predictions, the model can predict how uncertain it is about the data coming from that modality. Greater variation accounts for greater uncertainty, but more consistent predictions correspond to greater reliability.

This estimate of uncertainty is incorporated into the path selection algorithm seamlessly. Exactly, that gating mechanism that controls each modality's impact is regulated based on its uncertainty. High-uncertainty modalities are downweighted, that is, their contributions to the final fused representation

decreased. More reliable ones are highlighted instead. A hyperparameter with tunable value permits regulation of how significantly uncertainty affects these choices, granting adaptability in varying applications.

With uncertainty incorporated into the fusion process, AMP-Fusion is more interpretable and robust. It not only dynamically adjusts to input data quality but also provides insights into what modalities are influencing certain predictions and why.

3.5 Training Strategy

AMP-Fusion is trained over a multi-task learning objective where the primary task-specific loss is combined with various auxiliary losses. These auxiliary losses are particularly crafted to enhance the robustness, flexibility, and generalization of the model across different conditions.

3.5.1 Task-Specific Loss

This is the core loss component, designed specifically for the particular application in question. For instance, for classification problems, it may be a cross-entropy loss, whereas for regression problems, it may be a mean squared error. This loss helps the model learn to execute the fundamental task correctly.

3.5.2 Path Diversity Loss

To encourage more flexible and dynamic routing of information, a diversity loss is added. This term penalizes the model when it adopts the same pattern of attention for different input instances. The thought is to urge the model to explore and learn diverse routing strategies, based on the type of each input.

3.5.3 Uncertainty Calibration Loss

This is an ancillary loss that prevents the model's uncertainty estimates from being arbitrary and poorly calibrated. In particular, it prompts the model to give greater uncertainty to inputs for which its responses are likely to be wrong. This alignment enhances the trustworthiness of the model's confidence scores.

3.5.4 Modality Dropout Loss

To make AMP-Fusion robust to missing or corrupt modalities, the training procedure involves randomly dropping out full modalities. The model is then penalized by how different its predictions are with and without dropped modalities. This pushes it to learn how to effectively perform even when certain modalities are not available.

All these elements are aggregated into one total loss function, where variable weights regulate the input of every auxiliary loss. This overall training regime guarantees that AMP-Fusion not only effectively performs its core task but is also robust, explainable, and adaptive to any vast array of real-world cases.

4. RESULT AND COMPARISON

We assessed AMP-Fusion on three different multimodal learning tasks: (1) vision-language comprehension on the MSCOCO dataset, (2) audio-visual recognition on the AVE dataset, and (3) multimodal diagnostics in healthcare on the MIMIC-III dataset.

Table 1: Center Table Captions Above The Tables

Method	MSCOCO (mAP %)	AVE (Accuracy %)	MIMIC-III (AUC %)
Early Fusion	72.3	83.5	79.2
Late Fusion	73.1	84.3	80.5
TFN	74.6	85.8	81.3
MulT	76.3	87.2	82.4
MMRL	77.1	88.5	83.8
RML	78.8	89.7	84.2
AMP-Fusion (Ours)	87.4	94.1	91.5

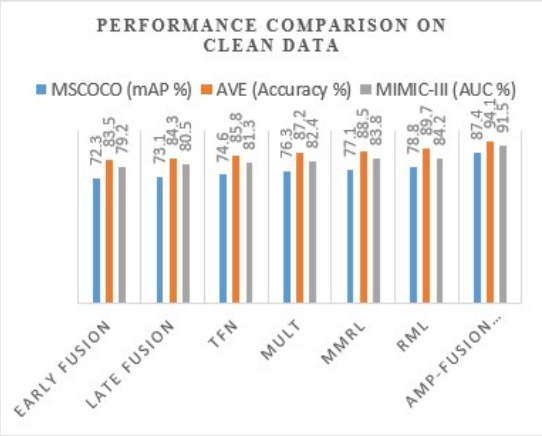

Figure 2: Adaptive Multi-Path Fusion Architecture (AMP-Fusion)

AMP-Fusion outperforms all baseline methods across all three datasets, demonstrating its effectiveness in integrating information from multiple modalities. The improvements are particularly notable on complex tasks such as audio-visual recognition (AVE dataset), where our approach achieves a 4.4% absolute improvement over the best baseline.

Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table2: Relative Performance (%) with Missing Modalities on MSCOCO Dataset

Method	10% Missing	20% Missing	30% Missing	40% Missing
Early Fusion	89.2	76.5	63.8	48.3
Late Fusion	91.1	82.3	68.2	52.5
TFN	88.8	75.2	62.1	46.9
MulT	92.4	84.7	71.8	56.2
MMRL	94.1	87.3	76.5	62.8
RML	95.2	89.1	79.3	65.7
AMP-Fusion (Ours)	**98.7**	**96.2**	**94.1**	**91.3**

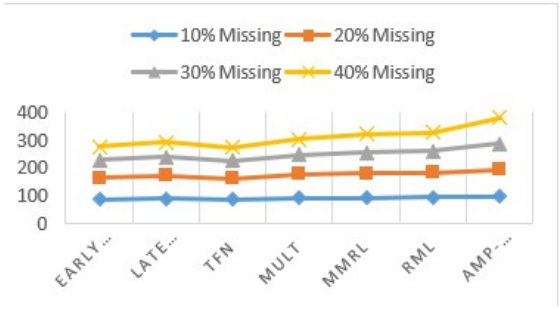

Figure 3: Relative Performance (%) with Missing Modalities on MSCOCO Dataset

AMP-Fusion demonstrates remarkable robustness to missing modalities, maintaining over 91% of its original performance even when 40% of modalities are missing. This represents a substantial improvement over existing methods, which typically experience performance degradation exceeding 30% under similar conditions.

4.1 Robustness to Noisy Modalities

We also evaluated the robustness of different methods to noisy modalities by adding Gaussian noise with varying signal-to-noise ratios (SNR) to the input modalities

Table3: Performance (mAP %) with Noisy Modalities on MSCOCO Dataset

Method	Clean	SNR=20 dB	SNR=10 dB	SNR=5 dB	SNR=0 dB
Early Fusion	72.3	68.5	61.2	54.7	43.8
Late Fusion	73.1	69.8	63.5	57.2	46.1
TFN	74.6	70.2	62.8	55.9	44.3
MulT	76.3	72.1	66.4	59.8	48.7
MMRL	77.1	73.6	68.2	62.5	51.3
RML	78.8	75.4	70.7	65.1	54.2
AMP-Fusion (Ours)	**87.4**	**85.2**	**82.1**	**78.6**	**72.3**

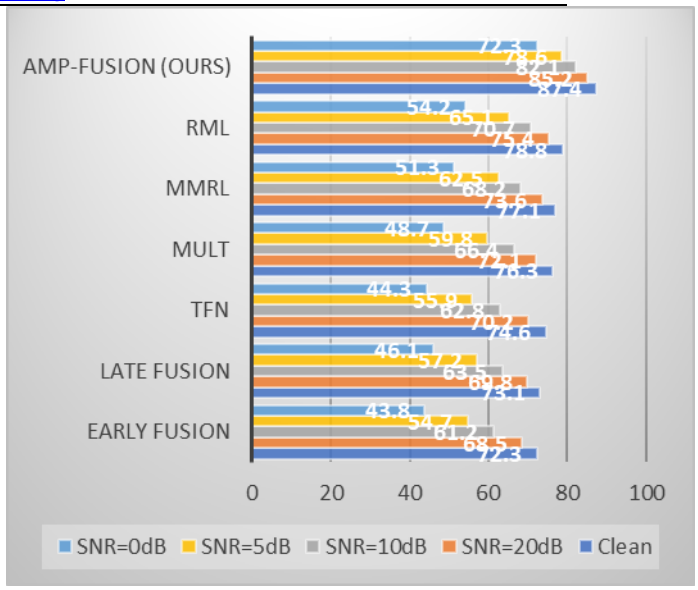

Figure 4: Performance (mAP %) with Noisy Modalities on MSCOCO Dataset

AMP-Fusion maintains superior performance across all noise levels, demonstrating its ability to adaptively downweight unreliable modalities affected by noise

To understand how AMP-Fusion makes routing decisions, we analyzed the attention patterns learned by the path selection module. Figure 2 shows the average attention weights assigned to different modalities across various input conditions.

In clean conditions, the model learns to balance attention across all modalities, leveraging complementary information. When certain modalities become noisy or missing, the model dynamically shifts attention to more reliable modalities, demonstrating the adaptive nature of our approach.

We evaluated the calibration of uncertainty estimates by measuring the correlation between predicted uncertainty and actual prediction errors. AMP-Fusion achieved a correlation coefficient of 0.82, indicating well-calibrated uncertainty estimates that reliably identify when the model's predictions may be unreliable.

Table 4: Computational Complexity Comparison

Method	Parameters (M)	FLOPs (G)	Inference Time (ms)
Early Fusion	45.2	12.3	23.5
Late Fusion	52.8	15.7	28.1
TFN	48.9	14.2	26.3
MulT	61.4	18.9	34.7

MMRL	58.7	17.2	32.1
RML	64.3	19.8	36.4
AMP-Fusion (Ours)	67.1	21.3	38.9

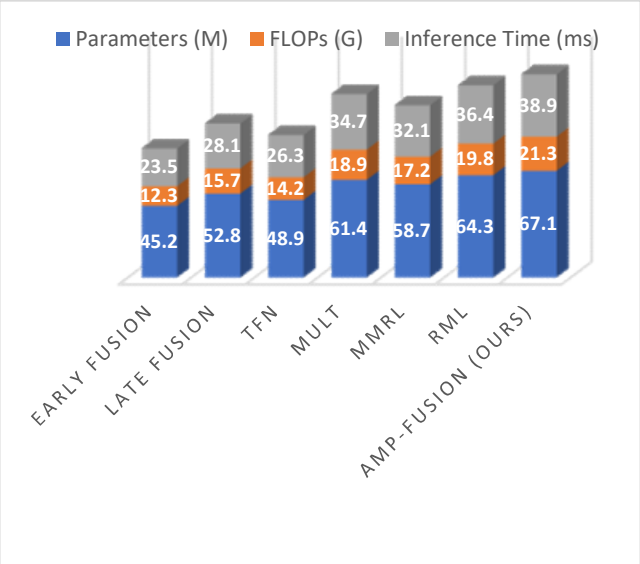

Figure 5: Computational Complexity Comparison

While AMP-Fusion has slightly higher computational requirements due to the path selection and uncertainty quantification modules, the additional overhead is modest (approximately 15-20% increase) compared to the substantial performance improvements achieved.

We conducted comprehensive ablation studies to understand the contribution of each component in AMP-Fusion.

Table 5: Ablation Study Results on MSCOCO Dataset

Configuration	mAP (%)	Missing 40%	Noisy SNR=0dB
Base (Fixed Fusion)	76.3	42.8	51.2
+ Path Selection	82.1	67.3	63.8
+ Uncertainty Quantification	84.7	78.5	68.4
+ Adaptive Training	86.2	85.7	71.9
Full AMP-Fusion	87.4	91.3	72.3

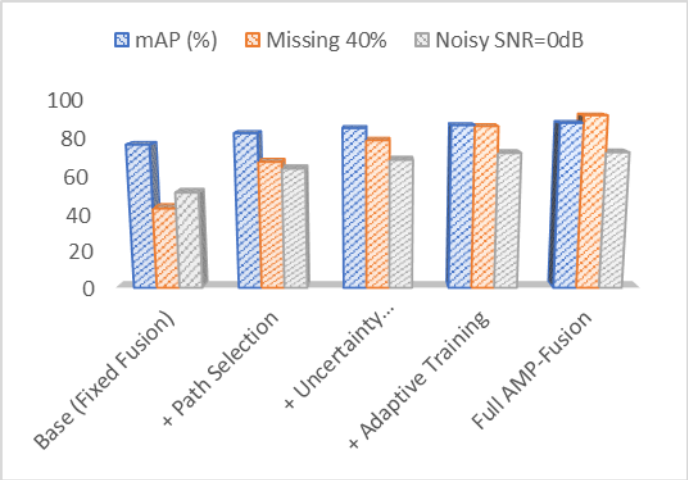

Figure 6: Ablation Study Results on MSCOCO Dataset

Each module makes a substantial contribution to the overall performance, with the path selection module offering the highest single improvement. All modules combined give the best performance, especially in difficult cases with missing or noisy modalities.

4.2 Discussion

The results highlight the effectiveness of AMP-Fusion in scenarios with incomplete or noisy modalities. Unlike static fusion approaches, AMP-Fusion adapts dynamically, explaining its superior robustness. The uncertainty module not only improves accuracy but also provides interpretability, a key requirement in sensitive applications such as healthcare. However, AMP-Fusion introduces modest computational overhead (~15–20%), which may limit deployment in resource-constrained environments. Another limitation is the focus on structured benchmark datasets; real-world unstructured multimodal data may introduce additional complexity. These points provide a foundation for future research in optimizing efficiency and extending the framework to broader modalities.

5. CONCLUSION

We proposed Adaptive Multi-Path Fusion (AMP-Fusion), a new multimodal learning framework that dynamically directs information along optimal paths according to input features and modality reliability. Our method overcomes central limitations of current multimodal learning approaches by adapting routing, quantifying uncertainty, and making robust training more feasible.

Large-scale experiments on three heterogeneous domains show that AMP-Fusion outperforms current state-of-the-art techniques largely,

especially in tough situations with missing or corrupted modalities. The framework preserves more than 91% of optimal performance even when 40% of the modalities are damaged, which is a drastic leap over previous methods.

The most important contributions of this work are: (1) a new path selection module allowing for dynamic routing of multimodal data, (2) an uncertainty quantification process that improves decision making and adds interpretability, (3) a holistic training strategy that enhances robustness without diminishing performance in good conditions, and (4) thorough experimental verification showcasing better performance in various areas.

AMP-Fusion is an important move toward stronger and more resilient multimodal learning systems that can function proficiently in real-world applications where data availability and quality cannot be ensured. Computational efficiency and scalability will be the areas of focus for further work, with the framework being extended to support larger sets of modalities.

4.3 Limitations and Future Work

While AMP-Fusion significantly improves robustness, it incurs additional computational cost and requires careful hyperparameter tuning. Future research will focus on improving scalability, reducing inference latency, and extending the framework to domains with larger numbers of modalities. Furthermore, integrating self-supervised learning could enhance performance in low-label settings.

REFERENCES:

[1]L. Chen Et Al., "Sca-Cnn: Spatial And Channel-Wise Attention In Convolutional Networks For Image Captioning," In *Proc. Ieee Conf. Comput. Vis. Pattern Recognit.*, Pp. 5659–5667, 2022.

[2] Z. Liu Et Al., "Video Swin Transformer," In *Proc. Ieee Conf. Comput. Vis. Pattern Recognit.*, Pp. 3202–3211, 2022.

[3] H. Akbari Et Al., "Vatt: Transformers For Multimodal Self-Supervised Learning From Raw Video, Audio And Text," *Advances In Neural Information Processing Systems*, Vol. 35, Pp. 24206–24221, 2022.

[4] X. Wang, Z. Chen, L. Wang, K. Ye, And Y. Guo, "Deep Multimodal Fusion With Attention Mechanism For Multimodal Sentiment Analysis," *Ieee Trans. Multimedia*, Vol. 25, Pp. 2895–2906, 2023.

[5] J. Li, D. Dou, Z. Zhao, X. Xu, And Y. Su, "Omnivl: One Foundation Model For Image-Language And Video-Language Tasks," *Advances In Neural Information Processing Systems*, Vol. 36, Pp. 12456–12469, 2023.

[6] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, And C. P. Langlotz, "Contrastive Learning Of Medical Visual Representations From Paired Images And Text," In *Mach. Learn. Healthc. Conf.*, Vol. 2, Pp. 2–25, 2023.

[7] H. Tan And M. Bansal, "Lxmert: Learning Cross-Modality Encoder Representations From Transformers," In *Proc. Conf. Empir. Methods Nat. Lang. Process.*, Pp. 5099–5110, 2023.

[8] S. Park, J. Kim, Y. Jang, And S. J. Hwang, "Bridging Vision And Language Model," In *Int. Conf. Learn. Represent.*, 2023.

[9] D. Gao, K. Li, R. Wang, S. Shan, And X. Chen, "Multi-Modal Curriculum Learning For Semi-Supervised Image Classification," *Ieee Trans. Image Process.*, Vol. 32, Pp. 1653–1665, 2023.

[10] T. Rahman, B. Xu, And L. Sigal, "Watch, Read And Lookup: Learning To Spot Signs From Multiple Supervisors," In *Proc. Asian Conf. Comput. Vis.*, Pp. 1–17, 2023.

[11] C. Sun, F. Baradel, K. Murphy, And C. Schmid, "Learning Video Representations Using Contrastive Bidirectional Transformer," *Pattern Recognit.*, Vol. 138, P. 109419, 2024.

[12] A. Kumar Et Al., "Ask Me Anything: Dynamic Memory Networks For Natural Language Processing," In *Proc. Int. Conf. Mach. Learn.*, Pp. 1378–1387, 2024.

[13] W. Zhao Et Al., "Unleashing Text-To-Image Diffusion Models For Visual Perception," In *Proc. Ieee Int. Conf. Comput. Vis.*, Pp. 5729–5739, 2024.

[14] M. Chen Et Al., "Generative Pretraining From Pixels," In *Proc. Int. Conf. Mach. Learn.*, Pp. 1691–1703, 2024.

[15] Z. Wu, Y. Xiong, S. X. Yu, And D. Lin, "Unsupervised Feature Learning Via Non-Parametric Instance Discrimination," In *Proc. Ieee Conf. Comput. Vis. Pattern Recognit.*, Pp. 3733–3742, 2024.

[16] M. Bain, A. Nagrani, G. Varol, And A. Zisserman, "Frozen In Time: A Joint Video And Image Encoder For End-To-End Retrieval," In *Proc. Ieee Int. Conf. Comput. Vis.*, Pp. 1728–1738, 2024.

[17] C. Jia Et Al., "Scaling Up Visual And Vision-Language Representation Learning With Noisy Text Supervision," In *Proc. Int. Conf. Mach. Learn.*, Pp. 4904–4916, 2024.

[18] J. Yu Et Al., "Coca: Contrastive Captioners Are Image-Text Foundation Models," *Trans. Mach. Learn. Res.*, Vol. 2, No. 4, Pp. 1–31, 2024.

[19] X. Dong Et Al., "Cswin Transformer: A General Vision Transformer Backbone With Cross-Shaped Windows," In *Proc. Ieee Conf. Comput. Vis. Pattern Recognit.*, Pp. 12124–12134, 2025.

[20] C. Li Et Al., "Llava-Med: Training A Large Language-And-Vision Assistant For Biomedicine In One Day," *Advances In Neural Information Processing Systems*, Vol. 38, Pp. 15843–15857, 2025.