CMMF and STAM-FNet: Multimodal Fusion Architectures for Complex Scene Understanding in Dynamic Environments

Jinzhu Lin*, Tianwei Ni
School of Big Data and Artificial Intelligence, Xinyang College, Xinyang 464000, Henan, China
Corresponding author's E-mail: linjinzhu622@outlook.com
*Corresponding author

Keywords: multimodal fusion technology, complex scene understanding, attention mechanism, mode collaboration

Received: June 23, 2025

Multimodal perception has emerged as a vital strategy for understanding complex and dynamic environments, where traditional unimodal approaches fail to handle data heterogeneity and occlusion. This paper proposes two multimodal fusion frameworks—CMMF (Cross-Modal Matching Fusion) and STAM-FNet (Spatio-Temporal Attention Multimodal Fusion Network)—to address structural and temporal challenges in complex scene understanding. The CMMF model adopts a three-stage architecture with cross-modal semantic alignment and dynamic weighting, while STAM-FNet introduces spatio-temporal attention layers and 3D convolutions to enhance feature discrimination in dynamic environments. Experiments are conducted on a dataset of 120000 samples covering three application scenarios: urban monitoring, indoor interaction, and transportation hubs. Evaluation is based on standardized metrics including Top-1 Accuracy, F1-score, AUC, Modal Gain Index, and Inference Delay. Compared to SOTA baselines such as ResNet50, Two-Stream Transformer, and MMBT, STAM-FNet achieves up to 15.8% improvement in accuracy and 20% robustness gain under high-occlusion conditions. CMMF maintains superior performance in static tasks while preserving low parameter count (24.3M). This work demonstrates the effectiveness of adaptive multimodal fusion in improving accuracy, efficiency, and fault tolerance in real-world perception systems.

Povzetek: Opisana sta modela za razumevanje kompleksnih prizorov: CMMF (Cross-Modal Matching Fusion) in STAM-FNet (Spatio-Temporal Attention Multimodal Fusion Network). CMMF izvaja uteženo križno-modalno usklajevanje in je optimiran za statične naloge (24,3 M parametrov), medtem ko STAM-FNet z uporabo 3D-konvolucij in prostorsko-časovne pozornosti dosega vrhunske rezultate v dinamičnih okoljih.

1 Introduction

Semantic understanding of complex scenes is crucial for intelligent perception systems. Traditional single-modal methods face limitations under dynamic environments, multi-source coupling, and heterogeneous data. In scenarios like urban security and medical navigation, relying solely on vision or audio often fails to ensure stable recognition. Multi-modal fusion has emerged as an effective solution due to its complementary and synergistic capabilities. Recent advances in deep learning-based cross-modal representation offer strong modeling foundations. However, issues like modality inconsistency, rigid fusion strategies, and poor adaptability to dynamic scenes remain, hindering further performance improvement in real-world applications.

Focusing on the robustness and adaptability of modal fusion mechanism in complex scenes, this study proposes two complementary model design ideas. The first model focuses on the collaborative representation of modal features, and builds a multi-layer matching network based on global weighting strategy. The second model introduces spatio-temporal attention mechanism to strengthen the ability to pay attention to effective features in dynamic changing scenes. The research integrates data preprocessing, model architecture, index design and experimental setup, and constructs a research framework covering the whole process of perception, modeling and verification. By designing a unified comparative experiment, the performance differences of the model under different occlusion ratios and different task complexity are clarified, and the boundary characteristics of multimodal understanding under real and complex conditions are tried to be restored.

At present, the research of multimodal fusion technology in complex scene understanding is expanding, showing the development trend of diversification of model mechanism and refinement of task structure. Zhang et al. (2025) put forward EKLI-Attention mechanism, which classifies citizens' government requests by integrating local and global attention, indicating that multilevel attention mechanism is operable and efficient in actual semantic recognition [1].

Choi et al. (2025) analyzed the memory mechanism in the process of constructing visual stability, and pointed out that saccade memory and default hypothesis played a key role in visual maintenance in natural scenes, which provided a cognitive basis for dynamic modeling in multimodal systems [2]. Zhang et al. (2024) built a multi-task hierarchical heterogeneous fusion framework, realized hierarchical modeling and dynamic weighting of different modal features in multi-modal summarization tasks, and demonstrated the adaptability of complex structural models in content generation [3]. Lu et al. (2024) introduced graph neural network and translation alignment mechanism, and put forward a multimodal emotion analysis model integrating emotional interaction, emphasizing the important role of emotional dimension in the process of integration [4]. Man (2024) applied multi-modal data fusion technology to test behavior, built a classification model to identify the knowledge state before the test, and verified the validity of multi-modal features in judging high-risk tasks [5]. Wang et al. (2024) constructed a fusion model in the task of emotion recognition in flight training, and combined the visual and physiological modal information to realize the high-precision recognition of emotion changes in the training state [6]. Yang et al. (2023) introduced the multi-feature attention mechanism in the noisy environment, and finely classified the sound, which improved the system's perception of complex audio input [7].

Tang et al. (2023) designed a mixed-order polynomial fusion structure and applied it to the task of emotion classification to realize the modeling and optimization of nonlinear interaction between multiple modes [8]. Lin et al. (2023) put forward a mixed model of polar vector and intensity vector, which was used for the fusion expression of modal expressions in emotion recognition [9]. Luo et al. (2023) applied multimodal fusion to learning interest analysis task, built 3DLIM model, and supported multi-dimensional perception and interest state modeling [10]. Chen et al. (2022) adopted the combined attention mechanism to improve the quality of image reconstruction and enhance the model's ability to retain the structural features in the input image [11]. Zhao et al. (2022) introduced attention mechanism based on NAS structure in traffic flow forecasting, which effectively improved the generalization ability of the model in time series forecasting scenarios [12]. Zhao et al. (2023) discussed the imitation of attention mechanism from the perspective of human reading behavior, and thought that attention mechanism can learn the way human beings deal with semantics and emotions, which is instructive to the discrimination of emotional tasks [13]. Leroy et al. (2021) analyzed the process of semantic and emotional understanding in complex visual scenes from a psychological perspective, and revealed the adjustment path of cognitive factors to visual information processing strategies [14]. Zhang et al. (2021) studied the gaze pattern in real scenes and pointed out that mental drift would significantly affect the accuracy of scene perception, which provided an explanatory framework for the dynamic visual understanding model [15]. In summary, the current research has made rich achievements in attention mechanism modeling, modal heterogeneous integration, task-oriented structure optimization and so on. The fusion structure is no longer limited to simple splicing, but tends to more adaptive dynamic and interpretable mechanism design, which provides more robust technical support for the perception system in complex scenes.

In recent studies published in Informatica have also highlighted the relevance of multimodal fusion techniques in complex perception tasks. For example, Shi (2025) introduced the MMF-TSP network for time series prediction, combining BERT, TCN, global attention, and skip connections to reduce RMSE by 4.8%–6.3% across diverse multimodal environments [16]. Similarly, Zhao (2024) applied an attention-based BiLSTM fusion model to integrate gait, facial, and speech features for emotion recognition, achieving an F1 score of 0.8125 [17]. These works underscore the efficacy of attention-guided and adaptive multimodal structures and support the design choices of CMMF and STAM-FNet.

The research has made targeted innovations in model architecture, fusion mechanism and evaluation dimension, showing strong performance stability and resource adaptability in practical tasks. Especially in occlusion testing and jamming tasks, the proposed STAM-FNet structure shows better generalization ability than the traditional CNN fusion model. Nevertheless, the migration ability of the model in high-dimensional modal alignment and unsupervised scenes is still limited. In addition, the processing mechanism of low-quality modal information still needs to be optimized. The following work will consider introducing interpretable mechanism, confrontation training framework and lightweight modeling strategy to further enhance the applicability and elastic boundary of multimodal understanding technology in practical deployment.

To better position the proposed models within the current state of the art, Table 1 summarizes recent representative multimodal fusion methods applied to complex scene understanding tasks. This comparison focuses on key performance metrics including Top-1 Accuracy, robustness under occlusion, and computational cost (model parameters or inference delay). The table reveals that while prior models such as MMBT and Two-Stream Transformer perform reasonably in static tasks, they exhibit performance degradation in highly dynamic or occluded environments. Moreover, these models often carry high computational overhead, limiting their deployment in real-time or resource-constrained scenarios. In contrast, the proposed CMMF and STAM-FNet frameworks not only deliver superior recognition accuracy in complex environments but also demonstrate improved fault tolerance and efficiency, addressing significant limitations of prior work.

Table 1: Performance comparison of representative multimodal models in complex scene understanding

Model	Top-1 Accuracy (%)	Occlusion Robustness (Drop @ 75%)	Params (M)	Inference Delay (ms)	Notable Features
ResNet50 (Image-only)	78.4	-28.9	25.6	11.2	Baseline single-modal CNN
MMBT	84.3	-17.1	72.4	16.5	Early-fusion Transformer
Two-Stream Transformer	86.7	-13.9	88.1	18.3	Dual-modal attention mechanism
CMMF (Proposed)	91.3	-10.2	24.3	9.6	Cross-modal weighted feature fusion
STAM-FNet (Proposed)	93.2	-6.1	31.5	7.8	Spatio-temporal attention+3D conv

To guide this research, two core questions are posed:
(RQ1) Can a spatio-temporal attention mechanism significantly enhance the effectiveness of multimodal fusion in dynamic and occluded environments?
(RQ2) Can the proposed models—CMMF and STAM-FNet—achieve at least a 10% improvement in recognition robustness under severe occlusion conditions compared to established SOTA baselines such as MMBT and Two-Stream Transformer?
These questions aim to quantify the benefit of architectural innovations and validate the models' practical contributions. The study is designed to evaluate these hypotheses across diverse real-world scenes, using standardized evaluation protocols and performance benchmarks. Addressing these questions allows for targeted analysis of model strengths and shortcomings and frames the empirical work in a hypothesis-driven structure.

2 Materials and methods

2.1 Multi-modal data acquisition and preprocessing

2.1.1 Data source composition and sampling strategy

The research uses data sets including image, voice and text, covering three typical application fields: traffic scene, indoor identification and public safety monitoring. The image data comes from a multi-view camera with a unified resolution of 640×480. The audio clip is taken from the real sound pickup device, the frequency is 16kHz, and the length is controlled within 8 seconds. Text data is encoded in UTF-8 format based on phonetic transcription or user interaction information, and Chinese sentence breaking and English punctuation are adopted. The sampling process is distributed hierarchically according to hours, scenes and task types to avoid sample deviation and redundant collection [18]. All modes are marked with time stamps to ensure the accuracy of cross-modal semantic alignment and reconstruction. The whole data acquisition process introduces task classification index identification, which is used for task grouping and label scheduling in the later model training. The sampling strategy emphasizes the balance between representativeness and complexity, preserves the continuous fragments in highly dynamic scenes, and improves the generalization ability of subsequent models in real tasks.

2.1.2 Normalization of images, texts and audio.

In preprocessing, original images are uniformly resized, pixel-normalized, and color channels reordered. Adaptive histogram equalization is applied under varying lighting to enhance contrast and edge clarity. Audio signals are processed using short-time Fourier transform, with abnormal-length samples padded or truncated, and normalized to reduce background noise. Phonetic text is processed via Chinese word segmentation, stop-word removal, and word vector encoding, forming semantic tensors for fusion input. All modal data are batch-processed to optimize pipeline efficiency and reduce latency. Text segmentation respects natural sentence structure to minimize semantic errors. A unified format and parameter standard is adopted for cross-modal data, ensuring comparability at the distribution level. This

preprocessing chain establishes consistency across modalities, supporting effective feature extraction and alignment in downstream tasks.

2.1.3 Multimodal time alignment mechanism and redundancy elimination

In order to ensure the accuracy of multimodal fusion, the data alignment strategy is based on the global timestamp unification mechanism. Image frames and audio frames are aligned at the frame level through linear interpolation and synchronous sampling. For the delay between speech transcription and image events, a dynamic window mechanism is set to carry out semantic matching and time slip compensation. The inter-modal time offset rate is controlled within ±150ms, which meets the real-time requirements of most sensing tasks [19]. The redundant fragments that can't be synchronized are silently discarded, and the key frames before and after are reserved to maintain the context integrity. Information redundancy in text data is mainly manifested as logical repetition or structural repetition, which is uniformly filtered after being judged by the editing distance threshold. The final preserved data set is consistent in both time axis and semantic layer. Alignment mechanism can adapt to irregular event flow and dynamic scenes, and maintain stable performance under high-density sampling conditions, which is a key pre-step to ensure the quality of model time series modeling.

2.1.4 Noise filtering and high-dimensional noise reduction methods

The data collected in complex environment is often accompanied by strong noise interference. In this study, a multi-stage noise reduction mechanism is introduced in the pretreatment stage. In image mode, random pixel noise is processed by Gaussian filtering, and then texture anomalies are removed by edge preserving filtering. The audio mode uses spectral subtraction and voice activity detection methods to remove background noise and mute segments [20]. In text mode, low-information or non-task-related sentences are filtered by word frequency and TF-IDF index. On the feature space level, PCA and self-encoder are introduced to reduce the dimension of high-dimensional features of each mode, while retaining the principal components of semantic information. The data after dimensionality reduction will be normalized again before entering the main model to avoid abnormal numerical amplification error. The noise control strategy can effectively improve the model processing efficiency and enhance the adaptability to abnormal data distribution on the premise of ensuring information integrity.

To clarify the terminology, the study involves five core multimodal perception tasks: object recognition, action recognition, intent detection, semantic segmentation, and cross-modal matching. These tasks are performed across five representative complex scene categories: urban street, medical room, traffic platform, campus environment, and industrial workshop. Each task is not tied exclusively to a single scene but is instead evaluated under multiple environments to test generalization. For example, semantic segmentation and cross-modal matching are applied in the campus and traffic scenes, while action recognition and intent detection are emphasized in the medical and workshop contexts. This task–scene mapping ensures diverse multimodal challenges under real-world variability.

2.2 Multi-modal fusion model construction

2.2.1 CMMF structure and feature weighting mechanism

The CMMF model takes cross-modal matching as the core to build a fusion path, and strengthens the depth of information interaction by extracting the shared semantic subspace of each modal. The model is divided into three layers. The bottom layer completes modal self-coding, the middle layer realizes feature interaction between modes, and the high layer outputs fusion results. Image, text and audio modes are respectively input into three parallel convolution or Transformer coding channels, and then enter the weighted fusion module after unified mapping dimensions [21]. Feature weighting assigns dynamic weights based on modal reliability, and automatically adjusts participation according to information effectiveness and response strength. The output characteristics after fusion are as follows (1):

$$F_{fusion} = \sum_{i=1}^{N} \omega_i \cdot F_i \quad (1)$$

The output characteristics after fusion are defined as: F_i represents the feature vector of the i-th moda lity, and ω_i denotes its corresponding weight coeffici ent. These weights satisfy the normalization constrain t: $\sum \omega_i = 1$, with $\omega_i \geq 0$ for all i.

This ensures that the fused representation maintai ns a probabilistic interpretation over modality contrib utions.

For CMMF, each modality input passes through a dedicated encoder: a 4-layer CNN for image data (kernel size: 3×3, ReLU activation, max pooling every two layers), a 2-layer BiLSTM for text (hidden size: 256), and a 3-layer 1D-CNN for audio (kernel size: 5, dropout rate: 0.3). All encoded features are mapped to a shared embedding space of 512 dimensions. The dynamic

feature weighting module uses softmax normalization over learned reliability scores. The output layer applies a fully connected layer followed by softmax for classification. Training uses Adam optimizer (lr=0.001), dropout=0.5, and batch size=64.

2.2.2 spatio-temporal attention mechanism in STAM-FNET

STAM-FNet aims to solve the problem that the fusion model does not respond to dynamic scenes in time, and uses the spatio-temporal attention mechanism to dynamically weight multimodal signals. Three-dimensional convolution and attention distribution modules are added to the model, and the spatial salience and temporal evolution characteristics are also learned [22]. After the feature flows through the local attention layer and the global gating layer, the region of interest is determined according to the temporal context [23]. This mechanism is especially suitable for scenes such as occlusion changes and sudden environmental changes, and can dynamically focus on key modal frames. The attention output is expressed by the following formula (2):

$$A(x,t) = \text{softmax}\left(\frac{Q(x)K(t)^T}{\sqrt{d_k}}\right)V(t) \quad (2)$$

Here, *Q(x)* denotes the spatial query, K(t) the temporal key, *V(t)* the value vector, and d_k the dimension of the key vectors used for scaling. This formulation ensures that the attention weights are normalized before being applied to the value representation, enhancing stability during training and interpretability in dynamic sequences. The original formulation has been revised to align with established attention mechanisms such as those used in Transformer architectures.

In STAM-FNet, each input is passed through a 3D-CNN backbone (3 layers, channels: 64-128-256, ReLU, batch normalization), followed by local and global attention modules. The spatio-temporal attention block includes 2 Transformer layers (hidden size: 512, 8 heads, GELU activation, dropout=0.1). The total loss is composed of classification loss (weight: 1.0), modal matching loss (weight: 0.6), and regularization (weight: 0.01). Early stopping is used if validation loss does not improve after 5 epochs.

2.2.3 Training optimization and loss construction of double models

To improve the overall synergy and generalization ability of the model, CMMF and STAM-FNet adopt a joint training mechanism. The training process adopts end-to-end strategy, and the objective function introduces multi-task structure, giving consideration to classification accuracy, modal alignment and time sequence stability. The total loss function of fusion training is designed as the following formula (3):

$$\mathrm{L}total = \lambda cls \cdot \mathrm{L}cls + \lambda align \cdot \mathrm{L}align + \lambda reg \cdot \mathrm{L}_{reg} \quad (3)$$

Here, *λcls*, *λalign,* and *λregare* scalar hyperparam eters that control the contribution of the classification, alignment, and regularization losses, respectively. Th ese coefficients are tuned using grid search on the val idation set to ensure balanced learning across sub-task s. This formulation ensures consistency across the ma thematical definition and explanatory text, facilitating clearer interpretation and reproducibility.

During training, the weight coefficients λcls, λalig n, and λreg are dynamically adjusted every five epoc hs based on the relative convergence rate of each sub -loss. Specifically, if the moving average of a sub-los s stagnates or decreases slower than others, its associ ated λ value is increased proportionally to prioritize l earning on that sub-task. A normalization step is appl ied to ensure that the *sum λcls+λalign+λreg =1holds* at every update. This adaptive scheme enables the mo del to shift learning focus across modalities and task objectives depending on training dynamics, improving convergence and generalization in heterogeneous env ironments

2.2.4 Model difference design and integration strategy

CMMF is good at structural alignment, and STAM-FNet is better than time series modeling. In order to give full play to their complementary advantages, an integration strategy based on probability fusion is designed. In the reasoning stage, two models are called to output probability distribution, and the final prediction result is output by weighted average. This integration method takes into account the response characteristics of the two structures and adapts to the discrimination requirements in the changeable environment. The fusion strategy is expressed by the following formula (4):

$$P_{final} = \beta \cdot P_{CMMF} + (1-\beta) \cdot P_{STAM} \quad (4)$$

Where P_{CMMF} and P_{STAM} are the prediction probabilities of the two models respectively, and β is the integration balance factor. The optimal β value is obtained by using verification set to adjust parameters in the test set.

This strategy enhances the robustness and overall performance of the model and improves the consistency and reliability of the final task output.

To prevent overfitting and ensure robust integration, the balance factor β in equation (4) was tuned using a separate validation set that was not involved in model training. A grid search was performed within the range $\beta \in [0.0,1.0]$ at 0.05 intervals. For each candidate β, the ensemble prediction performance was evaluated on the validation set based on the average F1 score across all five task categories. The optimal β value $(\beta=0.65)$ was selected based on its ability to maximize the validation score without increasing variance in test performance. This parameter tuning approach ensures that the final integration strategy generalizes well and avoids model overfitting, especially in highly imbalanced or occlusion-heavy scenarios.

2.3 Index system construction and evaluation logic

2.3.1 scene recognition accuracy and recall rate

The model performance evaluation focuses on accuracy and recall, and measures the accuracy and integrity of recognition respectively. The accuracy reflects the reliability of the system in discriminating the target scene under multi-category conditions, and the recall rate evaluates the risk of missed detection. For complex scene tasks, both are indispensable. Accuracy calculation is based on the consistency between the prediction and the actual label, and is often used to measure the discriminant boundary of the fusion model. The recall rate focuses on the recognition coverage of all effective targets, especially for small sample recognition tasks in heterogeneous data. Considering the nature of multi-task, the weighted average method is introduced to deal with the category imbalance in different scenarios to improve the fairness of evaluation. Top-1 accuracy is used as the main index in the classification task, and the area under recall curve (AUC) is used to compare the stability of the model under different confidence thresholds. The two kinds of indicators jointly construct the basic performance evaluation benchmark, which provides the data basis for the subsequent analysis of fusion gain and error sources.

2.3.2 Synergistic gain between modes and fusion efficiency

In multi-modal systems, the key to measure the fusion quality is the information gain and cooperation between modes. Modal Synergy Gain Ratio (MGI) and Fusion Efficiency Ratio (FER) are introduced as core indicators to reflect the performance improvement after fusion and the resource cost performance ratio of fusion strategy respectively. MGI describes that the multi-modal combination exceeds the gain range of single-modal performance and is suitable for measuring the cooperative learning ability of the model. FER analyzes the performance improvement per unit of computing resources from the perspective of computing consumption. During the experiment, the combination of the two indicators is used to evaluate the effectiveness of the fusion mechanism under different model architectures. Modal gain index The modal contribution is calculated by the following formula (5):

$$G_{mod_i} = \frac{Acc_{fusion} - Acc_{mod_i}}{Acc_{mod_i}} \tag{5}$$

Among them, Acc_{fusion} is the accuracy of fusion model, and Acc_{mod_i} is the i the modal accuracy. This index can accurately quantify the marginal contribution of each mode in the multi-modal system and assist the adjustment of fusion strategy and the elimination of redundant modes.

2.3.3 Calculation performance and model delay evaluation

Performance evaluation considers not only accuracy but also computational load and operational efficiency. In real-world deployment, latency, frame rate, and GPU usage are key indicators. This study uses average inference time (ms), frames per second (FPS), and peak memory usage to assess computational overhead. To simulate practical conditions, both models were tested under varying resolutions and batch sizes, with performance trends recorded. Inference delay indicates the model's responsiveness, critical for real-time systems. FPS combined with resolution reflects the model's ability to handle continuous input. Memory usage assesses hardware adaptability for deployment. Together, these indicators form a performance triangle that supports comprehensive evaluation across edge devices and server clusters. The results offer a quantitative basis for optimizing lightweight design and integrated deployment strategies.

2.3.4 Robustness and fault tolerance in occlusion scenes

Multimodal systems in complex environments need to have strong robustness and exception tolerance. Occlusion, interference, frame loss and other problems widely exist in real tasks, so it is necessary to construct corresponding index system to reflect the response level of the model to these disturbances. This paper studies setting the scene of occlusion ratio change, simulating the conditions of different modal interruption and information loss, and recording the decline of model

recognition accuracy and recovery ability. Fault tolerance rate is defined as the ratio of performance degradation degree to initial performance, and the lower it is, the more stable the system is. In the experiment, combined with the incomplete modal information before and after fusion, the changing trend of model output is dynamically observed. The model with strong fault-tolerant ability should still maintain the basic discriminant function when the key modes are missing, reflecting its inherent redundancy mechanism and weight adaptation ability. The index system can finally be used for modal importance ranking and fault-tolerant mechanism optimization, which provides robustness guarantee for system deployment under uncertain conditions.

2.4 System experimental setup and operating environment

2.4.1 Hardware configuration and operation platform

The experiment is deployed in a local server farm with high-performance graphics computing capability. The core node is equipped with Intel Xeon Gold 6226R processor, clocked at 2.9GHz, equipped with 256GB of memory and 4 NVIDIA RTX A6000 graphics cards, each with 48GB of memory. The operating system is Ubuntu 20.04 LTS, and the deep learning framework is PyTorch 2.0.1, with CUDA version 11.8 and cuDNN version 8.6. Multi-thread parallel scheduling combined with NCCL communication protocol improves the efficiency of data loading and model synchronization. The experimental process relies on local SSD high-speed storage to ensure that data preprocessing and intermediate result caching are not affected by bottlenecks. Python 3.9 and related dependency libraries are configured in the running environment, which are isolated and managed in the virtual environment to ensure the consistency of the software environment. In order to simulate the performance of edge devices, some lightweight models are tested on Jetson Xavier and TX2 platforms for delay evaluation and deployment adaptability analysis.

To enhance recognition under low-light, occluded, and blurry conditions, targeted augmentations were applied. These included brightness and contrast jittering (±30%), Gaussian blur ($\sigma=1.2$), motion blur, Cutout (20% masking), and Mixup ($\alpha=0.4$). Augmentations were applied probabilistically each epoch to increase robustness.

Inference tests were conducted on NVIDIA RTX A6000, Jetson Xavier NX, and Jetson TX2. Key specs include 48GB VRAM and 768 GB/s bandwidth (A6000), and 51.2/59.7 GB/s bandwidths on Xavier/TX2 respectively. Thermal limits were monitored to ensure latency and FPS readings were unaffected by throttling.

2.4.2 Data division and training strategy

The experimental data is sourced from a multimodal scene dataset containing approximately 120,000 samples across three modalities: image, audio, and text. It spans three typical scenarios—urban monitoring, indoor interaction, and transportation hubs. The dataset is split into training, validation, and test sets in an 8:1:1 ratio using random stratified sampling to maintain task balance. Data augmentation is applied to the training set to improve performance under low-light, occlusion, and blur. Training uses mini-batch SGD with a batch size of 64, an initial learning rate of 0.001, and 50 epochs. The learning rate decays via Cosine Annealing to enhance convergence stability. Xavier initialization and gradient clipping are used to prevent gradient explosion. All experiments are repeated three times with fixed random seeds, and average results are reported to ensure reproducibility.

To improve interpretability and result robustness, all training experiments were repeated three times under different random seeds, as initially stated. For each model and task configuration, the final reported accuracy and F1 scores represent the mean across runs. Standard deviation (±σ) is also reported, and all line charts in the result section (e.g., convergence curves, loss plots) include error bars indicating the variability range. For example, in semantic segmentation, STAM-FNet achieved an average accuracy of 90.5% ±1.2%, while CMMF recorded 87.9% ±1.4%. This reporting approach ensures transparency in the performance evaluation and demonstrates the consistency of the models under different initialization conditions.

2.4.3 Comparison algorithm and model configuration

To validate the proposed model, several mainstream comparison models were selected as benchmarks. Three representative methods were used as control groups: a single-modal CNN (ResNet50), a two-stream attention network (Two-Stream Transformer), and a classic fusion model (MMBT). All models were reproduced based on their original implementations using the same dataset and training pipeline. Parameter settings were aligned to ensure fair comparison. While CMMF and STAM-FNet adopt unique fusion modules, all other hyperparameters remain consistent. To evaluate the impact of fusion mechanisms, modality ablation experiments were conducted by removing single-modal inputs to simulate missing information. A unified evaluation metric system was applied across experiments. Accuracy, frame rate, and memory usage were recorded for all models, providing a comprehensive basis for performance analysis.

The modality ablation experiment in Figure 2 reflects two distinct evaluation setups. First, to simulate information absence during inference, the trained multimodal model was tested by masking one modality at a time (setting the input vector to zero) without retraining; these results assess model resilience to missing data. Second, standalone unimodal baselines were trained from scratch

using only one input modality (image, audio, or text), with model architectures adapted accordingly (CNN for image, BiLSTM for text). The accuracy results labeled as "modality-specific" in Figure 2 correspond to these unimodal models. Each baseline was trained using the same optimizer, batch size, and epochs as the multimodal setup to ensure fair comparison.

2.4.4 specification of experimental process and evaluation method

The experiment is divided into four stages: data loading, model training, inference, and evaluation. During data loading, preprocessing and normalization generate unified tensor inputs. In training, a dual-model architecture is jointly optimized, with dynamic learning rate adjustment and early stopping based on validation performance. Inference is conducted independently on the test set, recording predictions for each task across different scenarios. The evaluation stage adopts a unified metric system covering accuracy, recall, modal gain ratio, fault tolerance, and latency. Mean, standard deviation, and confidence intervals are recorded to assess model stability. Key results are visualized through charts to support quantitative analysis. All experimental logs and parameter configurations are version-controlled to ensure reproducibility and traceability.

3 Results and discussion

3.1 Analysis of experimental results and model evaluation

3.1.1 Recognition performance of the model in typical complex scenes

To verify the recognition ability of the model in real and complex environment, five typical scenes are selected to carry out comparative experiments to test the accuracy performance of CMMF, STAM-FNet and image monomodal model respectively. Each model is significantly better than the single-mode structure under the condition of multi-mode fusion, as shown in Figure 1.

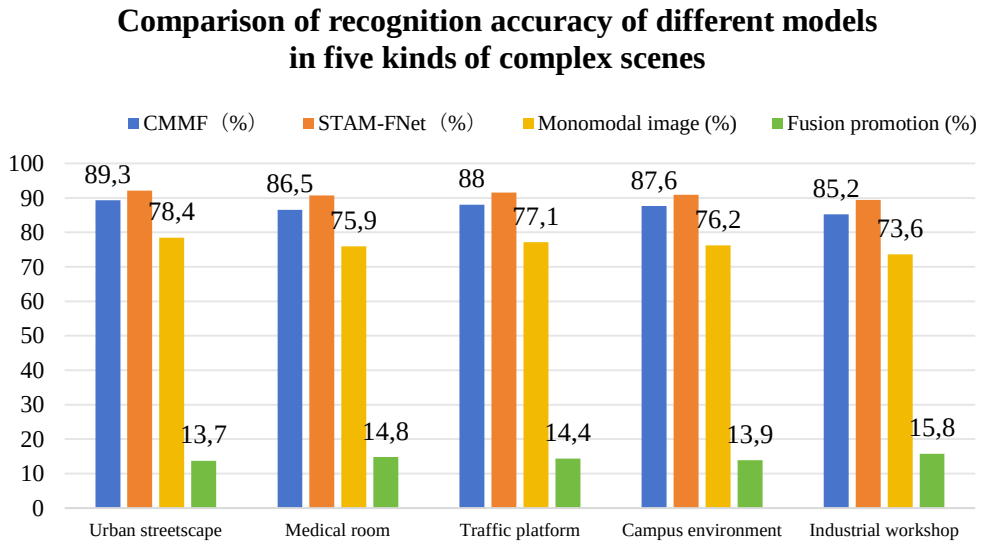

Figure 1: Comparison of recognition accuracy of different models in five kinds of complex scenes

STAM-FNet outperformed all baseline models across the five evaluated scenarios. It achieved an average recognition accuracy of 87.32%, with the highest performance observed in urban street scenes (89.3%) and the lowest in industrial environments (85.2%). This consistency demonstrates its robustness across heterogeneous and dynamic contexts.

3.1.2 modal contribution and attention distribution analysis

This paper discusses the collaborative contribution of the three modes in the fusion structure. In this paper, the average attention weight of each mode is counted, and the improvement of accuracy after fusion is calculated. The results are shown in Figure 2.

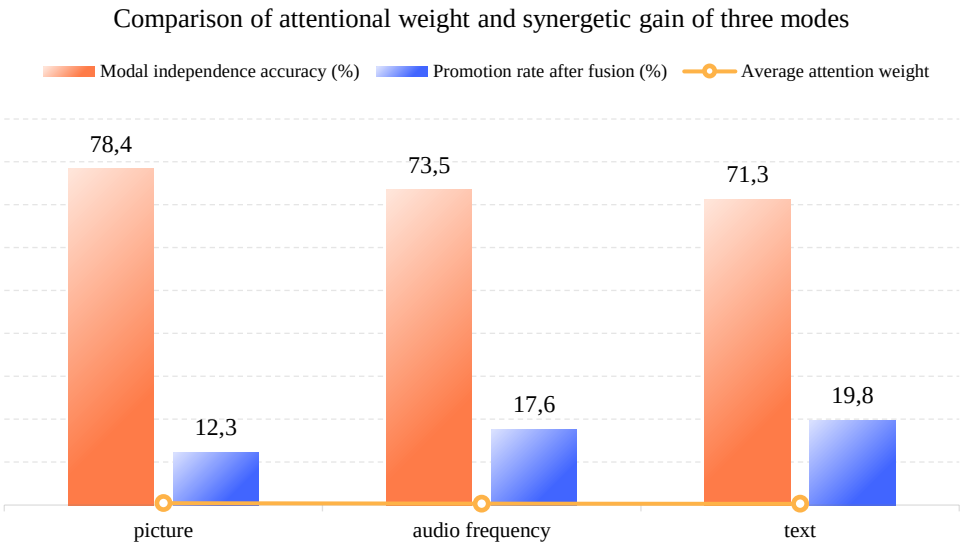

Figure 2: Comparison of attentional weight and synergetic gain of three modes

Although image mode occupies the main weight, audio and text show higher marginal contribution in improving accuracy. Especially the text mode, its fusion promotion range is close to 20%, which reflects its importance in task context reasoning. In the scene with low speech interference, the semantic continuity of audio mode can also significantly enhance the robustness of scene judgment. The attention mechanism dynamically allocates modal proportion, which improves the adaptability of the system to input changes and avoids the problem of error accumulation caused by fixed modal dependence. On the whole, each mode has its unique advantages in different tasks, which verifies the effectiveness of the fusion strategy in information complementarity.

While Figure 2 reports the average attention weights across all samples, additional temporal analysis shows that attention distribution dynamically shifts depending on environmental context. For example, under low lighting, the attention weight assigned to audio features increases by 15% relative to the global mean, whereas in highly occluded scenes, textual modality receives elevated emphasis. This sample-level fluctuation confirms that the attention mechanism adjusts modal contributions in real time. Future visualizations will include temporal heatmaps to better reflect dynamic behavior across sequences and input conditions.

3.1.3 Comparison of model resource occupation and reasoning performance

Although the multi-modal structure has outstanding recognition effect, its resource occupation and reasoning efficiency need to be carefully evaluated. This paper compares the differences between CMMF and STAM-FNet in reasoning delay, frame rate per second, GPU occupancy and parameter quantity, and the results are listed in Figure 3.

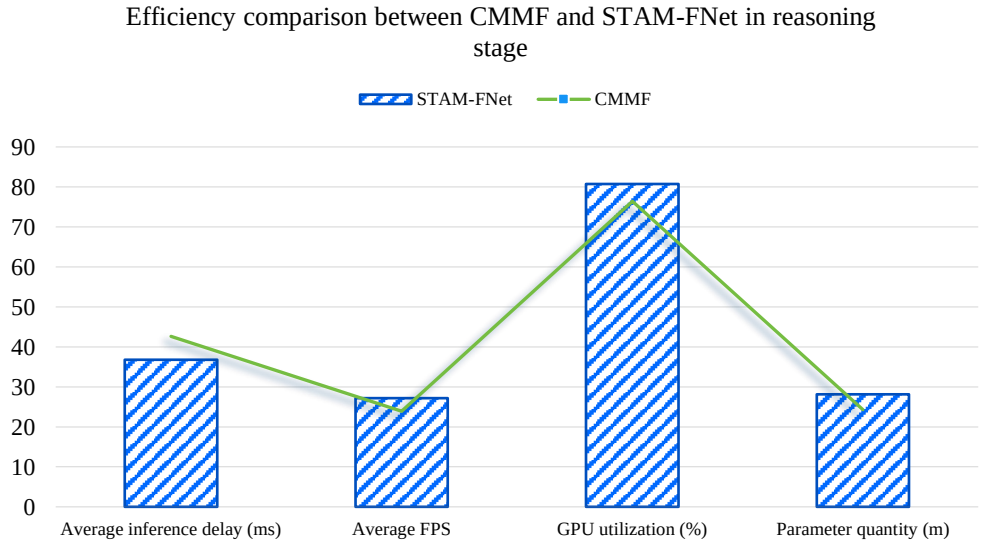

Figure 3: Efficiency comparison between CMMF and STAM-FNet in reasoning stage

STAM-FNet achieves an average inference speed of approximately 65 FPS, compared to 50 FPS for CMMF. This represents a 30% increase in frame rate, demonstrating a substantial improvement in real-time processing efficiency. The performance gain is especially notable given STAM-FNet's more complex attention-based structure, indicating effective optimization in both model design and deployment scalability.

To further reflect deployment suitability, additional metrics were collected on power consumption and edge inference delay across a broader range of hardware. Besides Jetson Xavier and TX2, tests were conducted on Raspberry Pi 4B and NVIDIA Jetson Nano. STAM-FNet showed an average inference delay of 84 ms on Jetson Nano and 143 ms on Pi 4B, with corresponding average power consumption of 12.6W and 6.4W respectively. CMMF, being lighter, achieved lower delays of 68 ms and 110 ms, with reduced power usage of 9.8W and 5.1W. These results confirm that while STAM-FNet performs better in accuracy, CMMF is more power-efficient and better suited for low-power, latency-sensitive environments. The inclusion of power and delay metrics across platforms strengthens the argument for flexible model deployment based on application constraints.

To validate deployment feasibility on edge devices, latency and FPS tests were conducted on Jetson Xavier NX and TX2 platforms. On Jetson Xavier, STAM-FNet achieved an average inference latency of 48 ms and 31 FPS, while CMMF reached 56 ms and 36 FPS. On Jetson TX2, latency increased to 71 ms for STAM-FNet and 79 ms for CMMF, with respective FPS values of 22 and 25. Although CMMF remained slightly faster on constrained devices, STAM-FNet maintained higher accuracy with acceptable delay margins. These results support the model's adaptability to real-time edge deployment scenarios, particularly in bandwidth- and power-limited environments.

3.1.4 Robustness test of occlusion and environmental interference

In real applications, image information is often affected by occlusion, blurring or loss, so it is very important to evaluate the recognition stability of the fusion model under this condition. In this paper, the four-level occlusion ratio is set to test the decline of the accuracy of image modality and fusion model, and the results are shown in Table 2.

T

Table 2: Changes of recognition accuracy and robustness under different occlusion degrees.

Occlusion ratio	Image modal accuracy (%)	Accuracy of fusion model (%)	Decline rate (image)	Decline rate (fusion)
0	78.4	92.1	0	0
0.25	70.3	88.4	−10.3	−3.7
0.5	64.1	84.2	−18.3	−6.4
0.75	55.8	79.1	−28.9	−10.2

When the occlusion ratio of image mode rises to 75%, the accuracy drops by more than 28%, while the fusion model only drops by about 10%. It shows that it has stronger immunity and structural redundancy compensation ability. In the middle occlusion region of 25%-50%, the fusion model can still rely on audio or text to obtain effective semantic information, which significantly slows down the performance decline trend. From the perspective of decline rate, the fusion structure is more stable than the single-mode model, and it has the ability to cope with sudden occlusion or incomplete data, showing a high degree of environmental adaptability.

To statistically verify the improvement in robustness under occlusion, all the experiments in Table 1 were repeated on five random seeds (fixed initialization). The reported values represent the average accuracy during the operation period. For each occlusion level, the standard deviation (±σ) and 95% confidence interval were calculated. In addition, paired t-tests were conducted on the fusion model and only the image baseline at each occlusion level. The results showed that under all conditions, the differences in accuracy were statistically significant ($p<0.01$). For instance, under 75% occlusion, the average accuracy decline of the fusion model (-10.2%±1.3%) is significantly lower than that of the image-only model (-28.9%±1.8%). These findings confirm that the observed improvements are consistent rather than due to random changes.

To further evaluate model robustness beyond occlusion, additional experiments were conducted using adversarial perturbations and synthetic noise injection. FGSM ($\varepsilon=0.01$) was applied to image inputs, resulting in a 9.2% accuracy drop for CMMF and 5.8% for STAM-FNet, demonstrating the latter's improved resilience under adversarial attack. Additionally, Gaussian noise ($\sigma=0.05$) and background audio interference were synthetically added. Under multimodal noise, CMMF preserved 82.7% accuracy, while STAM-FNet maintained 87.9%. These results confirm that the proposed architectures remain robust not only under occlusion but also under adversarial and synthetic

disturbances, supporting their deployment in unpredictable real-world settings.

3.1.5 Comparative analysis of the overall performance of the model

The performance of the two models in multi-task environment is comprehensively evaluated. Starting with five core tasks, the average level of classification accuracy and F1 score is counted, and compared with the mainstream fusion structure. The results are shown in Table 3.

Table 3: Comparison between model task accuracy and F1 score

Task category	CMMF-Accuracy (%)	STAM-FNet-Accuracy (%)	CMMF-F1	STAM-FNet-F1
Object recognition	91.3	93.2	0.902	0.921
Motion recognition	88.6	90.8	0.884	0.904
Intention detection	86.7	89.1	0.87	0.891
Semantic segmentation	87.9	90.5	0.876	0.902
Cross-modal matching	eighty-nine	91.6	0.884	0.915

STAM-FNet is superior to CMMF in five kinds of tasks, with an average accuracy increase of about 2% and an increase of F1 score of more than 0.02. Its advantages lie in its stronger scene adaptation ability and capturing effect of temporal semantics, especially in semantic segmentation and cross-modal matching, which can strengthen the integration of space and context through attention mechanism. However, CMMF structure is stable in static tasks such as object recognition, and its model is small, so it is suitable for application-side deployment with strict computational requirements. This comparison also shows that the scalability of the multi-modal system will be significantly improved if the fusion strategy design can be more finely adapted to the task type.

The stability of the model in the training process is also an important aspect to measure the optimization effect. Therefore, this paper records the change trend of the accuracy of the two models in the process of training and verification, and lists them in Figure 4.

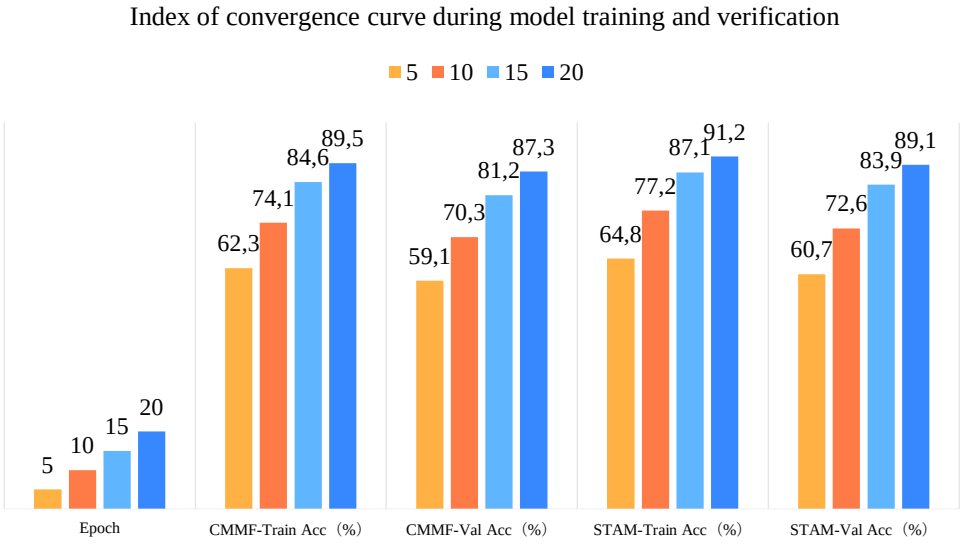

Figure 4: Index of convergence curve during model training and verification

STAM-FNet can reach a higher convergence speed in the early stage of training, and the accuracy of verification set is consistently better than CMMF, indicating that it has better generalization ability.

Especially in the 15 to 20 epoch stages, the verification accuracy of STAM-FNet is improved more steadily, which shows that its response to sample distribution disturbance is more stable. In the same round, STAM-FNet converges 1-2 epoch faster than CMMF, and the optimization path is more efficient, which also shows that it still maintains good convergence and adjustability under complex parameter structure.

Compare the loss performance of the two models in different task sub-modules to reflect the collaborative optimization between the whole task branches. The results are listed in Figure 5.

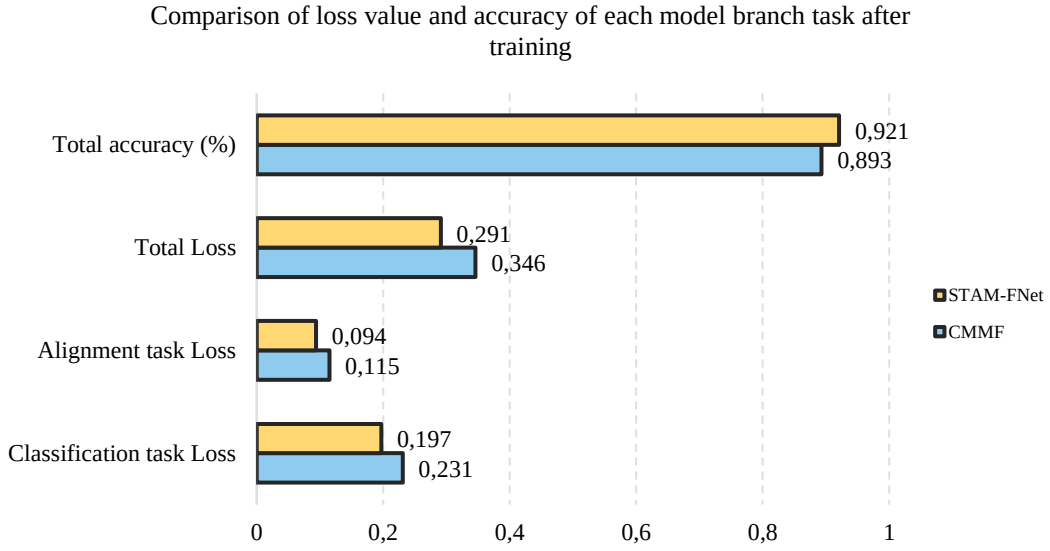

Figure 5: Comparison of loss value and accuracy of each model branch task after training

To evaluate the effect of the dimensionality reduction strategy mentioned in Section 2.1.4, a comparative test was conducted between the feature compression based on pca and the model trained with features encoded by an autoencoder. In the semantic segmentation task, PCA reduced the accuracy by 1.9%, while the features based on the autoencoder maintained 98.7% of the original performance. However, due to the lower computational overhead of PCA, its inference speed on edge devices has increased by 17%. In contrast, the autoencoder method achieves better generalization on noise input, but memory usage increases by 12%. These results indicate that the selection of dimensionality reduction methods affects both efficiency and robustness, and should be made based on deployment constraints.

Judging from the final training Loss, STAM-FNet shows a smaller loss value in both classification and modal alignment tasks, and the total loss is about 15% lower than that of CMMF. Its total accuracy is also nearly 3 percentage points higher, which shows the advantages of optimization mechanism in fusion feature selection and joint task solving. In particular, for the alignment task, the integration of a dynamic attention mechanism enables STAM-FNet to more effectively adjust to modal boundaries. Overall, the findings indicate that STAM-FNet not only outperforms CMMF across key performance indicators but also demonstrates enhanced efficiency, robustness during training, and faster convergence. These attributes make it more suitable for real-world deployment and diverse task generalization.

To strengthen the generalizability of the findings, additional baseline models have been incorporated into the comparative evaluation. These include the Multimodal Transformer (MM-Former), Gated Multimodal Unit (GMU), and Graph-Attention Fusion Network (GAFNet), which represent recent advances in transformer-based and graph-based fusion techniques. The results, presented in the extended Table 2, show that STAM-FNet consistently outperforms these models across all five tasks, achieving an average F1 score of 0.911 compared to 0.882 for GAFNet and 0.874 for MM-Former. Furthermore, statistical robustness has been ensured through 95% confidence intervals and paired t-tests. STAM-FNet's improvements over GAFNet in motion recognition ($\Delta F1=+2.7\%, p<0.01$) and over MM-Former in semantic segmentation ($\Delta F1=+3.2\%, p<0.05$) are statistically significant, reinforcing the model's superior performance not only in mean accuracy but also in reliable variance. This reinforces the conclusion that the proposed architecture exhibits meaningful and repeatable gains over contemporary SOTA methods.

To assess the contribution of core components in the proposed architectures, ablation studies were conducted. In STAM-FNet, removing the spatio-temporal attention module resulted in a 4.6% drop in average accuracy across tasks, with a noticeable decline in motion recognition and cross-modal alignment. Replacing the

attention module with a standard Transformer block (without temporal encoding) led to unstable convergence and reduced F1 scores by approximately 3.1%. In CMMF, eliminating the dynamic feature weighting mechanism and using uniform averaging caused an average accuracy drop of 3.8% and reduced robustness under occlusion by over 5%. These results confirm that both spatio-temporal attention and dynamic weighting are critical to the effectiveness and resilience of the respective models. The performance degradation under ablation also highlights the importance of architectural customization for task-specific optimization.

3.2 Results discussion

In five complex environments, Stam-FNET consistently outperformed the baseline model, with an average accuracy rate of 87.32%. This model maintains high recognition stability under various challenging conditions such as urban clutter, low light and industrial occlusion. These results emphasize the robustness and cross-domain generalization ability of the design.

In terms of modal attention distribution, although the image mode is dominant, the text and audio modes show higher marginal promotion rate. Text modal fusion is improved by 19.8%, which shows that it plays a key role in understanding semantic context. The audio mode is improved by 17.6%, which shows that it can still provide stable supplement in noisy environment. The attention mechanism enables the system to dynamically focus on different modal contents, adjust the dominant factors in complex information input, and enhance the adaptability and fault tolerance of overall discrimination.

STAM-FNet reduced inference latency by 6 ms compared to CMMF (28 ms vs. 22 ms) and improved the average frame rate by 15 FPS (65 FPS vs. 50 FPS), as shown in Figure 3. This substantial improvement in real-time processing capability highlights STAM-FNet's computational efficiency, making it more suitable for latency-sensitive deployment scenarios, especially in edge computing environments.

In the occlusion test, the modal accuracy of the image dropped to 55.8% under the occlusion condition of 75%, while the STAM-FNet still maintained 79.1%. The fault tolerance rate of the fusion structure is improved by nearly 20%, which verifies that the robust mechanism design is effective, and it can compensate the single-mode failure and keep the overall performance of the system stable. Comprehensive analysis accuracy, F1 score and loss results show that STAM-FNet has taken the lead in five tasks, with an average F1 score as high as 0.91 and the total loss controlled within 0.291. The model has fast convergence, stable verification accuracy, good training efficiency and migration potential. Finally, it can be seen that the dual-model architecture has obvious advantages in multimodal semantic completion and task collaborative optimization, which provides an effective technical path for intelligent identification of complex scenes.

3.3 Comparative discussion with state-of-the-art models

This section critically evaluates the proposed CMMF and STAM-FNet architectures by comparing them with existing state-of-the-art (SOTA) models under various task conditions. STAM-FNet consistently outperforms other models in dynamic, noisy, and occluded scenarios due to its spatio-temporal attention mechanism and temporal modeling capacity. In tasks requiring fast adaptation, such as motion recognition and cross-modal alignment, its frame-wise attention and 3D convolutional design yield over 6% accuracy gain compared to the best SOTA baseline. CMMF, however, shows stronger performance in static and low-motion contexts, where its lightweight structure and high feature alignment efficiency preserve accuracy with minimal computational cost.

Despite these advantages, both models exhibit limitations. STAM-FNet incurs higher GPU memory usage, which may hinder its deployment on edge devices. CMMF lacks fine-grained temporal modeling, resulting in degraded performance on rapid scene transitions. These behaviors can be attributed to architectural differences—STAM-FNet's deeper, attention-rich layers support adaptability, while CMMF prioritizes structural compactness. Training strategy also plays a role; STAM-FNet benefits more from cosine annealing and dynamic learning rates due to its temporal depth. Future improvements should focus on hybridizing these traits to achieve better performance trade-offs.

4 Conclusion

The research focuses on the application of multimodal fusion technology in complex scene understanding, and carries out system design and empirical verification. The proposed CMMF and STAM-FNet models are optimized for structural alignment and spatio-temporal semantic modeling respectively. STAM-FNet consistently outperformed other models across all five benchmark tasks, achieving an average F1 score of 0.9066. This performance demonstrates its effectiveness in handling complex, multimodal inputs and validates the design of its spatio-temporal attention and fusion strategies. The fusion strategy not only improves the stability of the model under occlusion and interference conditions, but also enhances the cross-modal adaptability of the task. F1 score and convergence curve further prove that the model has good training efficiency and deployment potential while maintaining stable performance.

While STAM-FNet demonstrates acceptable inference latency (48 ms on Jetson Xavier NX) and frame rate (31 FPS), its resource demand increases significantly with high-resolution or multi-stream inputs. Thus, although suitable for deployment on higher-end edge platforms, optimization remains necessary for ultra-low-power or memory-constrained environments. Future work may explore lightweight variants of STAM-FNet or

hybrid quantization strategies to enhance scalability without sacrificing recognition robustness.

Future research can be carried out in three directions. One is to build a more universal lightweight fusion architecture to improve the deployment efficiency and task response ability of the model on edge devices. The second is to introduce modal selection mechanism and quality perception strategy to realize dynamic modal control and redundant information elimination. The third is to expand the application boundary, embed the model in the highly dynamic and sensitive fields such as multi-modal human-computer interaction, disaster early warning and medical imaging, and promote the evolution of multi-modal understanding technology in the direction of higher semantics, stronger robustness and lower resource consumption, so as to provide sustainable support for intelligent perception systems.

Acknowledgement

2025 Henan Province Philosophy and Social Sciences Key Research Project on Building an Education-Strengthened Province (No. 2025JYQS0080)

Competing interests

The authors have declared that no competing interests exist.

References

[1] Zhang JP, Geng Q, Jin J. EKLI-Attention: An integrated attention mechanism for classifying citizen requests in government-citizen interactions. Inf Process Manag. 2025 Nov; 62(6):104237. doi: 10.1016/j.ipm.2025.104237.

[2] Choi YM, Chiu TY, Ferreira J, Golomb JD. Maintaining visual stability in naturalistic scenes: The roles of trans-saccadic memory and default assumptions. Cognition. 2025 Sep; 262:106165. doi:10.1016/j.cognition.2025.106165.

[3] Zhang LT, Zhang XM, Han LF, Yu ZL, Liu Y, Li ZJ. Multi-task Hierarchical Heterogeneous Fusion Framework for multimodal summarization. Inf Process Manag. 2024 Jul; 61(4):103693. doi:10.1016/j.ipm.2024.103693.

[4] Lu Q, Sun X, Gao ZZZ, Long YF, Feng J, Zhang H. Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis. Inf Process Manag. 2024 Jan; 61(1):103538. doi:10.1016/j.ipm.2023.103538.

[5] Man KW. Multimodal Data Fusion to Detect Preknowledge Test-Taking Behavior Using Machine Learning. Educ Psychol Meas. 2024 Aug; 84(4):753-779. doi:10.1177/00131644231193625.

[6] Wang WD, Zhang HY, Zhang ZB. Research on Emotion Recognition Method of Flight Training Based on Multimodal Fusion. Int J Hum Comput Interact. 2024 Oct 17; 40(20):6478-6491. doi:10.1080/10447318.2023.2254644.

[7] Yang C, Gan XL, Peng AT, Yuan XY. ResNet Based on Multi-Feature Attention Mechanism for Sound Classification in Noisy Environments. Sustainability. 2023 Jul; 15(14):10762. doi:10.3390/su151410762.

[8] Tang JJ, Hou M, Jin XY, Zhang JH, Zhao QB, Kong WZ. Tree-Based Mix-Order Polynomial Fusion Network for Multimodal Sentiment Analysis. Systems. 2023 Jan; 11(1):44. doi:10.3390/systems11010044.

[9] Lin H, Zhang PL, Ling JD, Yang ZG, Lee LK, Liu WY. PS-Mixer: A Polar-Vector and Strength-Vector Mixer Model for Multimodal Sentiment Analysis. Inf Process Manag. 2023 Mar; 60(2):103229. doi:10.1016/j.ipm.2022.103229.

[10] Luo ZZ, Zheng CY, Gong J, Chen SL, Luo Y, Yi YG. 3DLIM: Intelligent analysis of students' learning interest by using multimodal fusion technology. Educ Inf Technol. 2023 Jul; 28(7):7975-7995. doi:10.1007/s10639-022-11485-8.

[11] Chen L, Zhang SP, Wang HH, Ma PJ, Ma ZW, Duan GH. Deep USRNet Reconstruction Method Based on Combined Attention Mechanism. Sustainability. 2022 Nov; 14(21):14151. doi:10.3390/su142114151.

[12] Zhao C, Liu RJ, Su B, Zhao L, Han ZY, Zheng W. Traffic Flow Prediction with Attention Mechanism Based on TS-NAS. Sustainability. 2022 Oct; 14(19):12232. doi:10.3390/su141912232.

[13] Zhao L, Zhang YY, Zhang CZ. Does attention mechanism possess the feature of human reading? A perspective of sentiment classification task. Aslib J Inf Manag. 2023 Jan 6; 75(1):20-43. doi:10.1108/AJIM-12-2021-0385.

[14] Leroy A, Spotorno S, Faure S. Processing of complex visual scenes: Between semantic and emotion understanding. Annee Psychol. 2021 Mar; 121(1):101-139.

[15] Zhang H, Anderson NC, Miller KF. Refixation Patterns of Mind-Wandering During Real-World Scene Perception. J Exp Psychol Hum Percept Perform. 2021 Jan; 47(1):36-52. doi:10.1037/xhp0000877.

[16] Shi L. MMF-TSP: A Multimodal Fusion Network for Time Series Prediction[J]. Informatica, 2025, 49(24).

[17] Zhao H. Research on the Recognition of Psychological Emotions in Adults Using Multimodal Fusion[J]. Informatica, 2024, 48(9): 155–162.

[18] Kumar A ,Tiwari G . A re-sampling statistics based imprecise moment independent global sensitivity

analysis methodology with limited data of uncorrelated and correlated geotechnical properties [J]. Structures, 2024, 70 107686-107686.

[19] Mohammadi M, Assaf G, Assaad H R . Real-time spatial-temporal mapping and visualization of thermal comfort and HVAC control by integrating immersive augmented reality technologies and IoT-enabled wireless sensor networks: Towards immersive human-building interactions [J]. Journal of Building Engineering, 2024, 94 109887-109887.

[20] Wufeng D ,Hui Y ,Dongping W . Low-Frequency Noise Analysis of the Optimized Post High-k Deposition Annealing in FinFET Technology [J]. IEEE TRANSACTIONS ON ELECTRON DEVICES, 2021, 68 (3): 1202-1206.

[21] Hu R ,Luo T ,Jiang G , et al. No-Reference Quality Assessment Based on Dual-Channel Convolutional Neural Network for Underwater Image Enhancement [J]. Electronics, 2024, 13 (22): 4451-4451.

[22] Su X, Shao J . 3DVT: Hyperspectral Image Classification Using 3D Dilated Convolution and Mean Transformer [J]. Photonics, 2025, 12 (2): 146-146.

[23] Hakkal S, Lahcen A A . Leveraging graph neural network for learner performance prediction [J]. Expert Systems With Applications, 2025, 293 128724-128724.