European Journal of AI, Computing & Informatics

Vol. 1 No. 2 2025

Article **Open Access**

Design and Implementation of AI-Based Multi-Modal Video Content Processing

Da Xu [1,*]

[1] Video Infra, Meta, Menlo Park, CA, 94025, USA
* Correspondence: Da Xu, Video Infra, Meta, Menlo Park, CA, 94025, USA

Received: 01 June 2025
Revised: 08 June 2025
Accepted: 23 June 2025
Published: 25 June 2025

Abstract: Multimodal information interaction is gradually becoming an important direction for intelligent video content understanding. In videos, image, voice, and text collaboratively form a semantic system, which goes beyond the capabilities of single-modal information analysis. Efficient extraction and fusion of multi-source information has become a key challenge in artificial intelligence applications for various tasks such as classification, summarization, and content monitoring. Current research tends to focus on single-task or single-modal processing, and there is still a lack of universal fusion frameworks. In this context, establishing a universal, highly integrated, and well scalable AI multimodal video processing framework not only conforms to the trend of technological development, but also provides reliable technical support for intelligent communication, social services, educational innovation, and more.

Keywords: multimodal fusion; video comprehension; deep learning; artificial intelligence framework

1. Introduction

Multimodal information interaction is gradually becoming an important direction for intelligent video content understanding. Image, voice and text build semantic system together in video, which has gone beyond information analysis based on single mode. Efficient extraction and fusion of multi-source information has become a key challenge in artificial intelligence applications for various tasks such as classification, summarization, and content monitoring. Current research tends to focus on single task or single modal processing, and there is still a lack of universal fusion frameworks. In this context, establishing a universal, highly integrated, and well scalable AI multimodal video processing framework not only conforms to the trend of technological development, but also provides reliable technical support for intelligent transmission, public governance, social education reform, and more.

2. Overview of Multimodal Video Content Processing Technology

Multimodal video data refers to a type of artificial intelligence research direction that focuses on data fusion and intelligent analysis, which simultaneously contains multiple sources of information such as image information, sound information, and text information in a video. As a carrier of information from multiple sources, Video exhibits unique characteristics in each modality: images convey spatial features such as object positions and colors; sound captures language and emotions; text typically includes labels, subtitles, or scene descriptions [1]. Therefore, by understanding the mixed characteristics

of images, sound, and text, it is possible to deeply capture events, characters, emotions, and actions in videos, enabling the system to meet more needs such as intelligent recognition, search, and recommendation.

Nowadays, multimodal processing technology continues to break through the limitations of model architecture and fusion mechanisms, and various advanced fusion methods have emerged. For example, Transformer based multimodal models adopt attention mechanisms to achieve meaning alignment and complementarity between different modalities, which have excellent effects in video summarization, scene recognition, emotion detection, and other fields. The fusion method has gradually evolved from simple feature concatenation to hierarchical alignment, dynamic weighting, adaptive selection, and other methods, enabling the model to flexibly adjust the processing path based on differences in information types. The rapid evolution of the evolving frameworks of multimodal video content processing have promoted the application of artificial intelligence in multiple fields such as media, education, and public security [2].

3. Design of AI Based Multimodal Video Processing Framework

Multimodal video content processing is the process of model creation and information fusion across multiple sources such as images, sounds, and texts. To achieve intelligent semantic recognition and multitasking concurrency, the system needs to build an architecture with a clear structure, collaborative functions, and real-time deployment capability. By introducing the design of an AI based video processing framework, the system elaborates on its overall layout, core functional module division, and collaborative communication mechanism between modules, providing technical foundation and structural support for subsequent applications (Figure 1).

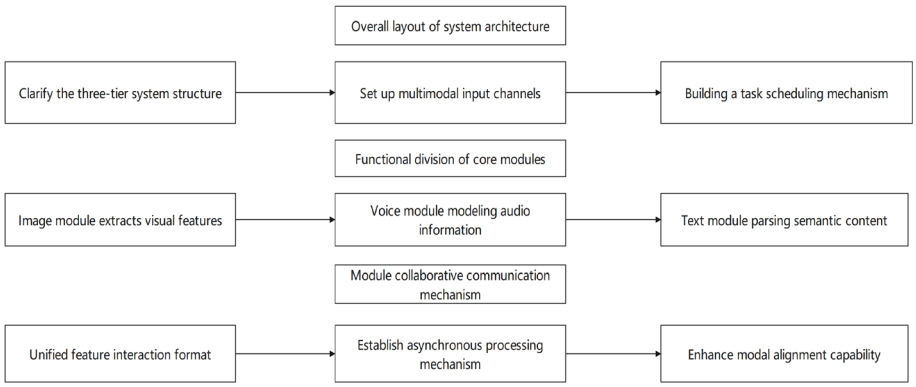

Figure 1. Design Based on AI Multimodal Video Processing Framework.

3.1. Overall Layout of System Architecture

The typical artificial intelligence-based multimodal video processing system framework consists of four parts: input, modal feature extraction, fusion decision, and task output. The system receives raw video data as input, and after decoding, it obtains image frame sequences, speech tracks, and embedded text data. Extracting spatial structural features through CNN, modeling temporal information using MFCC and RNN for audio, and generating word embedding representations using LSTM language model for text [3]. The various modal features are then uniformly fed into the multimodal fusion module. During the process, a modal attention mechanism is added to adaptively adjust the feature contribution of different modalities based on their importance. Let the image, speech, and text features be represented respectively; the fused representation can then be defined as:

$$\mathrm{F_{fusion}} = \alpha_{\mathrm{v}}\mathrm{F_v} + \alpha_{\mathrm{a}}\mathrm{F_a} + \alpha_{\mathrm{t}}\mathrm{F_t} \quad (1)$$

of which $\alpha_v + \alpha_a + \alpha_t = 1$, indicate the weight coefficients of each modality. This representation is further input into downstream task modules for completing functions

such as event recognition, content summarization, and tag generation. The overall architecture balances information integrity, computational efficiency, and scalability to meet the intelligent processing needs of multi scene videos.

3.2. Functional Division of Core Modules

Integrating image processing, speech processing, word processing, and inference into a multimodal video processing solution, the image module utilizes deep convolutional networks to extract spatial features from frame sequences, focusing on target actions, scene layout, and visual dynamics; The speech module characterizes and processes audio signals in the form of spectrograms, using bidirectional circulators to track speech rate, tone, and semantic information; The text module applies OCR to recognize text content in videos and trains language models on large datasets to obtain contextual semantic information [4]. Each module independently completes its corresponding tasks to ensure the targeted and accurate extraction of different data features, while the fusion module serves as a liaison function, fusing information from various modules and effectively aggregating information from different sources. This module helps to efficiently integrate various signals into one. The image features can be set as x_v; The speech features are x_a; The text features are x_t; The fusion formula is:

$$Z = ReLU(W_v x_v + W_a x_a + W_t x_t - b) \tag{2}$$

of which $W_v,\ W_a,\ W_t$ for the trainable weight matrix corresponding to the modality, b is the bias term, and ReLU is the activation function. The fused feature vector will serve as a unified input for downstream tasks such as classification and detection, supporting the system's semantic understanding ability in complex scenarios.

3.3. Module Collaborative Communication Mechanism

The multimodal video processing system consists of multiple independent sub modules, which require efficient communication mechanisms to achieve information sharing and functional collaboration. Due to inherent differences in sampling frequency, feature dimension, processing time, etc. among different modalities, an intermediate state buffer and a unified data transmission protocol were designed to eliminate these differences, while each module adopts asynchronous scheduling to dynamically allocate runnable resources through task publishing. At the same time, to prevent individual channels from becoming processing bottlenecks, the system has designed a delay adjustment method with multimodal perception capability [5]. Before integrating all features, communication methods need to follow standardized operating procedures and perform fusion processing on information from different sources. The characteristics of the three modalities of image, speech, and text are as follows $f_v,\ f_a,\ f_t$. The system generates shared intermediate expressions through a weighted summation mechanism:

$$\mathrm{f_c} = \frac{\mathrm{W_v f_v + W_a f_a + w_t f_t}}{\mathrm{W_v + W_a + W_t}} \tag{3}$$

of which $w_v,\ W_a,\ W_t$. The weight coefficients for each modality are dynamically determined by the attention score during the training process. This formula not only achieves the unification of feature dimensions, but also provides a stable foundation for subsequent semantic fusion, ensuring consistency and collaboration of information between different modalities during transmission.

4. Problems in AI Based Multimodal Video Processing Framework

4.1. Low Accuracy of Modal Data Fusion

In multimodal video processing, it is difficult to achieve high-precision matching in the fusion stage due to fundamental differences in data, semantic and temporal layers among different modalities such as image, sound, and language. Images focus on spatial vision, sound focuses on continuous time, and text points to the transmission of abstract

meaning. The three modalities exhibit heterogeneous characteristics and lack consistent structural alignment, making fusion in feature space prone to semantic conflicts or information imbalance. If there is a lack of dynamic adjustment algorithms, it will affect the accuracy of multimodal collaboration, leading to feature redundancy or loss of key information. Meanwhile, if the acquisition rates and processing delays among different modalities are inconsistent, it can cause temporal misalignment and lead to fusion failure. In some cases, modal loss, unstable quality, or noise interference may weaken the stability of the semantic system, thereby affecting the accuracy of multimodal understanding tasks.

4.2. Poor Deployment Efficiency of Deep Models

Typically, multimodal video processing systems utilize deep neural network structures to achieve various functions such as image recognition, speech recognition, and text generation, greatly increasing the size and computational complexity of the entire model, resulting in significant resource consumption in inference processing. This type of model has a strong dependency on GPUs or high-performance computing platforms, and may struggle to meet the requirements for real-time processing on edge devices or in low-resource environments. In addition, due to the separate operation of each mode channel and the lack of a unified compression mechanism, it increases the burden on the system and brings about extended time. When multiple tasks are executed concurrently, data processing and feature integration are prone to accumulate, increasing system pressure and easily forming computational bottlenecks, resulting in a decrease in overall work efficiency. Some frameworks currently lack effective adjustment methods, which can result in unstable performance and excessive memory usage when handling a large number of short video tasks, making it difficult to maintain system stability and scalability. Although deep learning models can effectively represent data, the inefficiency of their deployment remains a key challenge for implementing multimodal video processing systems.

4.3. Weak Adaptability of System Migration Capability

Multimodal video processing systems often struggle to maintain stable performance in different scenarios after training on specific tasks or datasets. Due to differences in semantic value and distribution across modalities in various tasks or applications, the ability of trained models to be applied in new fields is limited. These models are highly dependent on the type, quality, and structure of the input modalities. When modal loss, noise increase, mixing, and other situations occur, the operational ability will fluctuate dramatically. Some models lack flexible architecture adjustment capabilities and cannot make small adjustments or replace components in a timely manner according to specific environments, thereby reducing their deployment efficiency in practical applications. In addition, factors such as data protection, interface differences, and inconsistent device performance mostly reduce their applicability across platforms. Due to the significant differences in video content and task requirements, this system migration capability has become a technical bottleneck for its popularization and sustainable development.

5. Implementation Strategy Based on AI Multimodal Video Processing Framework

5.1. Improving the Accuracy Level of Cross Modal Fusion

To improve integration accuracy, a unified semantic description system adopts an automatic harmonization strategy to balance the information weights across different modalities. Multimodal video information, such as spatial visual information, emotional music, and semantic content, has significant differences in the types and methods of expression features for each mode. The multimodal attention structure enhances the perception of key information and suppresses the influence of invalid features or interfering modalities, thereby improving the overall expression effect. The encoding stage standardizes the features of each modality to provide a unified expression for subsequent fusion. The fusion stage adopts dynamic gating or weighting mechanisms to achieve effective alignment

and integration of multi-source data. The comprehensive description not only integrates semantic features across modalities but also enhances the system's adaptability in complex environments (Table 1).

Table 1. Cross Modal Feature Fusion Mechanism.

Modal type	Main features	Encoding method	Integration mechanism
Image modality	Spatial structural information	CNN/ResNet	Spatial attention fusion
Speech modality	Tone and rhythm	MFCC + LSTM	Time weighted mechanism
Text modality	Contextual semantics	BERT/RoBERTa	Multi head attention mechanism

When performing multimodal video evaluation tasks, screen images, background speech, and user text comments were simultaneously inputted. The system processes various modal features in parallel through a multi-channel encoding network and introduces a multi-head attention mechanism for dynamic matching during the fusion stage. If the characters in the image are heavily occluded or the audio signal is weak, the allocation ratio of the text mode will be automatically increased to ensure the correctness of the expression recognition results. Experiments have shown that this fusion strategy can maintain stable performance even in the presence of quality fluctuations in various modalities, with an accuracy improvement of over 11% in facial expression recognition compared to baseline methods, demonstrating enhanced robustness and cross-modal flexibility.

5.2. Design an Efficient and Deployable Network Architecture

The multimodal video processing framework faces challenges in practical implementation, such as limited computing resources, large model size, and diverse deployment platforms. Building an efficient and deployable network architecture requires synchronous optimization in three aspects: lightweight design, module decoupling, and platform adaptation. The use of depthwise separable convolution, attention pruning, knowledge distillation, and other methods can reduce redundancy and maintain stable performance. Using a shared encoder to extract features from images, speech, and text enables effective cross-modal interaction while preserving modality-specific information, allowing the modalities to learn from and reinforce each other. During the deployment phase, the model format and graphics optimization can be ensured through intermediate state converters and accelerators, thereby ensuring fast running speed and good stability on various hardware platforms (Table 2).

Table 2. Efficient and Deployable Network Optimization Techniques.

Optimize dimensions	Technical method	Deployment effect
model structure	Depthwise separable convolution and pruning	Reduce parameter count and memory usage
feature coding	Shared Encoder Design	Reduce duplicate channels and improve execution efficiency
Platform adaptation	TensorRT/ONNX compilation	Realize fast loading and inference across multiple platforms

In a mobile video detection task, network video recognition and speech recognition tasks need to be implemented on a low-energy hardware platform. On the basis of retaining the original multimodal model, lightweight processing was carried out, reducing the number of parameters by about 60%. Merging the video and audio streams into a shared encoding layer allows reuse of low-level features, thereby avoiding redundant computations. The model optimized based on TensorRT has increased the running speed by 1.8 times on the ARM architecture terminal platform, and the latency remains within 200ms. It can run stably on different models of intelligent terminals, effectively verifying the feasibility and generalization of high-speed and easy to implement designs in practical work.

5.3. Strengthen the Adaptability of Heterogeneous Environment Models

To adapt to the needs of various computing devices such as servers, mobile devices, edge devices, etc., a multimodal video processing system needs to have great scalability and flexibility in implementation, which can be achieved through module decoupling design, elastic structure, and automatic compilation optimization mechanism. At the model level, a loose coupling approach is adopted to separate the preprocessing, pattern encoding, fusion calculation, and output results of input information into independent modules. Different image processing pipelines are dynamically selected based on platform performance, such as choosing lightweight models for edge devices and deeper networks for cloud servers. In the fusion stage, a multi-path routing mechanism is introduced to extend the functionality of core computing nodes to meet performance requirements in various computing environments. At the deployment level, the model is transformed into an execution graph on a specified device using intermediate representation conversion tools (such as ONNX) and automatic graph optimization libraries (such as TVM), and is accelerated through compilation (Table 3).

Table 3. Heterogeneous Platform Model Adaptation Strategy.

Adaptation strategy	Applied technology	Adapt to the target platform
Module decoupling	Independent subgraph construction, submodule switching	Mobile/Edge Devices
dynamic trim	Channel pruning and substructure selection mechanism	Low to medium power processing unit
automatic optimization	ONNX export, TVM image compilation	GPU/ARM/FPGA platform

In the context of urban traffic monitoring projects, the system deploys view analysis models both in the cloud and at the edge. The separated architecture allows video and audio to be configured separately. In edge device deployment, the system automatically compiles lightweight model versions through TVM and uses the ONNX format for fast model loading. With the adoption of model quantization and compression techniques, the execution latency on embedded terminals is controlled within 150ms, and its memory usage is reduced to about half. In multi scenario testing, the system can flexibly adapt to different platform computing conditions, ensuring real-time performance and system stability in multimodal processing tasks.

6. Conclusion

This study focuses on an AI driven multimodal video processing framework, providing a practical and scalable solution from system design and core modules to fusion mechanisms and deployment strategies. By optimizing modal features, reducing model construction scale, and enhancing adaptability to heterogeneous environments, the multimodal framework can effectively handle complex video processing tasks with good accuracy. Demonstrating the framework's stability and generalizability, multimodal processing has proven effective in various applications including but not limited to text interpretation, facial analysis, and event detection. Multimodal fusion and intelligent computing are expected to drive innovation across diverse fields such as education, security surveillance, and digital media in the future, providing stronger support for video content understanding.

References

1. S. von Hertzberg-Boelch, P. M. P. Dworschak, D. Niemann, A. B. Verheyden, A. Bülhoff, M. M. Moche, et al., "An informational video for informed consent improves patient comprehension before total hip replacement—a randomized controlled trial," *Int. Orthop.*, vol. 49, no. 6, pp. 1303–1308, 2025, doi: 10.1007/s00264-025-06503-6.
2. S. Di Pietro, G. Tamburini, C. La Manna, R. Spagnolello, G. Romagnoli, S. Toccafondi, et al., "Video clips for patient comprehension of atrial fibrillation and deep vein thrombosis in emergency care. A randomised clinical trial," *NPJ Digit. Med.*, vol. 7, no. 1, p. 107, 2024, doi: 10.1038/s41746-024-01107-7.

3. V. Agrawal, M. V. V. Kantipudi, and J. Jagtap, "Enhancing hand-drawn diagram recognition through the integration of machine learning and deep learning techniques," *Sci. Rep.*, vol. 15, no. 1, p. 1, 2025, doi: 10.1038/s41598-025-01823-4.
4. J. Park, J. Lee, J. Choi, S. Kim, H. Yoon, K. Han, et al., "NEST-C: A deep learning compiler framework for heterogeneous computing systems with artificial intelligence accelerators," *ETRI J.*, vol. 46, no. 5, pp. 851–864, 2024, doi: 10.4218/etrij.2024-0139.
5. L. Cheng and X. Gong, "Appraising regulatory framework towards artificial general intelligence (AGI) under digital humanism," *Int. J. Digit. Law Gov.*, vol. 1, no. 2, pp. 269–312, 2024, doi: 10.1515/ijdlg-2024-0015.