Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 15 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i15.57101

PAPER

Optimized Fusion of Audio-Video Archival Data in Mobile Environments via the DeepSeek Multimodal Algorithm

Ruhua Bai (✉)

Luoyang Institute of Science and Technology, Luoyang, China

lylgbrh@lit.edu.cn

ABSTRACT

With the widespread adoption of mobile intelligent devices and the rapid development of 5G technology, audio-video archival data have been increasingly utilized in mobile scenarios, requiring more advanced multimodal fusion capabilities to support intelligent and mobile archival management. However, limitations in mobile device computing power, unstable network conditions, and challenges posed by the spatiotemporal asynchrony and semantic heterogeneity of audio and video data have rendered traditional fusion approaches inadequate in balancing accuracy with low-power, real-time processing on mobile platforms. Existing models based on conventional machine learning or early deep learning architectures have limited the ability to address the fragmentation and asynchrony of multimodal data in mobile environments. Therefore, a DeepSeek-based multimodal algorithm tailored for mobile contexts was proposed in this study. The approach focuses on the design of a soft-threshold attention module, the formulation of a polarized loss function, and the construction of a lightweight fusion network architecture. Dynamic cross-modal feature weighting was employed to optimize computational resource allocation, while the modified loss function enhanced semantic coupling. The network design was adapted to the constraints of mobile hardware to enable efficient and real-time fusion of multimodal data. The outcomes of this study provide a lightweight solution for processing complex audio-video archival data on mobile terminals, offering significant implications for improving archival resource utilization and promoting the integration of archival management with mobile computing technologies.

KEYWORDS

mobile environment, audio-video archival data fusion, DeepSeek multimodal algorithm, soft-threshold attention module, polarized loss function

Bai, R. H. (2025). Optimized Fusion of Audio-Video Archival Data in Mobile Environments via the DeepSeek Multimodal Algorithm. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(15), pp. 26–40. https://doi.org/10.3991/ijim.v19i15.57101

Article submitted 2025-04-10. Revision uploaded 2025-06-07. Final acceptance 2025-06-10.

1 INTRODUCTION

With the proliferation of mobile intelligent devices and the advancement of 5G technology [1–4], audio-video archival data have been increasingly applied in mobile scenarios such as law enforcement recording, remote conferencing, and emergency command. In such environments, these data exhibit inherently multimodal characteristics [5, 6], encompassing heterogeneous types such as audio semantics, video imagery, timestamps, and geolocation information. Efficient fusion of these modalities is critical to advancing the intelligent and mobile management of archival systems. However, mobile terminal devices are generally constrained by limited computational capacity, unstable network transmission, and insufficient battery life [7, 8]. Furthermore, the spatiotemporal asynchrony and complex semantic associations across modalities in audio-video data introduce significant challenges. Under these conditions, traditional data fusion methods have struggled to simultaneously satisfy the accuracy, real-time processing, and low-power requirements of mobile platforms, highlighting the urgent need for targeted optimization strategies to enhance multimodal data fusion efficiency.

Existing audio-video data fusion approaches have predominantly relied on conventional machine learning or early deep learning models [9, 10], such as shallow fusion networks based on feature concatenation or end-to-end models employing spatiotemporal attention mechanisms [11]. While these models have achieved some success in fixed computing environments, their limitations in mobile contexts have become increasingly apparent. On one hand, traditional models are characterized by large parameter counts and high computational complexity [12, 13], rendering them poorly suited for lightweight deployment on mobile devices. This mismatch often results in excessive energy consumption and high response latency during fusion processes. On the other hand, current methods inadequately address issues such as noise interference and data packet loss during mobile network transmission [14, 15] and suffer from limited precision in spatiotemporal feature alignment. Notably, deficiencies remain in capturing dynamic semantic coupling across modalities—for instance, in identifying correlations between spoken instructions and corresponding visual actions [16]. Moreover, most existing studies have not incorporated dedicated fusion strategies that account for the unique characteristics of archival metadata, making it difficult for the degree of structural coherence in the fusion outputs to meet the specialized requirements of archival management.

Focusing on the fusion of audio-video archival data in mobile environments, this study proposed a DeepSeek-based multimodal algorithm adapted for such scenarios. It explored three key components: First, a soft-threshold attention module was developed to dynamically adjust cross-modal feature weights, improving resource use on mobile devices. This enhanced spatiotemporal alignment while reducing computational load. Second, a polarized loss function was designed and compared to standard functions such as cross-entropy and triplet loss. It was optimized for semantic correlation and modality balance in archival data, improving the model's ability to handle heterogeneous inputs. Third, a lightweight fusion network based on attention mechanisms was built for real-time audio-video data processing on mobile hardware. The main innovation lies in moving away from dependence on fixed computing environments. By combining lightweight design with improved semantic representation across modalities, the study offers an efficient solution for processing complex archival data on mobile platforms. These findings improve the quality and efficiency of multimodal fusion in mobile settings and provide a new

direction for intelligent archival applications. They also offer useful theoretical and practical guidance for integrating archival management with mobile computing.

2 FUSION OF AUDIO-VIDEO ARCHIVAL DATA IN MOBILE ENVIRONMENTS

In mobile environments, the fusion of audio-video archival data is challenged by limited device computational capacity, unstable network transmission, and the heterogeneous nature of diverse data modalities. For instance, body-worn law enforcement recorders are required to integrate real-time audio, video, and geolocation data; in mobile office scenarios, smartphones must efficiently consolidate meeting audio, visual footage, and textual summaries; and in remote archival access, multimodal data must be synchronously processed under bandwidth constraints. The DeepSeek multimodal algorithm was designed with cross-modal feature alignment, a lightweight network architecture, and dynamic adaptive fusion capabilities. These characteristics enable the optimization of model parameters for mobile computing environments, thereby reducing power consumption without compromising fusion accuracy. A noise-resilient mechanism was also integrated to address issues such as packet loss and transmission delay in mobile networks. Through joint modeling of spatiotemporal features, the synchronization of audio and video data was significantly enhanced. In archival data processing tasks, the algorithm enables the precise extraction of deep associations between audio content, video frame features, and metadata.

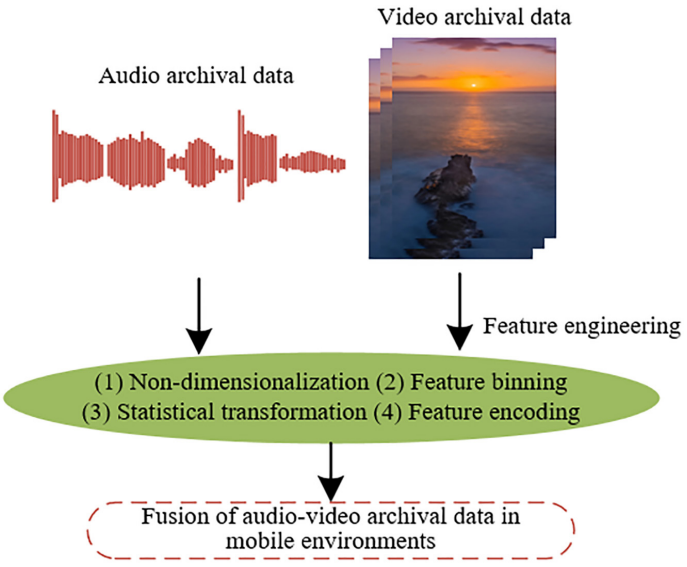


Fig. 1. Preprocessing workflow for audio-video archival data fusion

Practical applications include the accurate alignment of spoken commands with visual actions in surveillance footage and the structured fusion of audio-video data with textual transcripts in courtroom proceedings. These capabilities effectively mitigate the fragmentation and asynchrony typical of multimodal data in mobile scenarios, thereby offering robust technical support for the efficient utilization of audio-video archival data on mobile platforms. The preprocessing workflow prior to fusion is illustrated in Figure 1.

2.1 Design of the soft-threshold attention module

A soft-threshold attention module was introduced into the DeepSeek multimodal algorithm in response to the dual requirement for lightweight feature selection and adaptive computational allocation in mobile environments. In scenarios involving mobile law enforcement recorders or inspection devices, terminal units are required to process multimodal archival data—comprising audio, video, and geolocation information—in real time. However, the constrained central processing unit or graphics processing unit (CPU/GPU) resources of mobile platforms pose limitations, and the global weighted operations of conventional attention mechanisms often result in excessive energy consumption and latency, thus failing to meet real-time processing demands. The soft-threshold attention module addresses these challenges by dynamically generating threshold values for cross-modal feature weights. This enables the selective filtering of redundant information while preserving key semantic associations, thereby reducing computational complexity. For example, in noisy law enforcement scenarios, this module facilitates the precise alignment of spoken instructions with corresponding visual actions. In remote meeting recordings, it enables focused coupling between a speaker's voice and their close-up visual frames. The weight computation logic was optimized to accommodate mobile hardware characteristics. By applying a soft-thresholding function, adaptive pruning of feature weights was achieved. This approach not only avoids the information loss typically introduced by hard-thresholding but also ensures that attention computation remains within the processing limits of mobile platforms. As a result, the module effectively resolves the contradiction between "feature overload" and "computational bottlenecks" in mobile multimodal data, offering a robust solution for applications such as law enforcement archiving and mobile meeting summary generation through efficient cross-modal feature alignment.

The core principle of the soft-threshold attention module lies in constructing a dynamic weighting mechanism across modalities to achieve lightweight multimodal fusion in mobile contexts. In terms of input-output relationships, the module accepts audio semantic features, visual features from video data, and associated archival metadata. These inputs are encoded into unified cross-modal feature vectors through a feature encoding layer. The soft-threshold function was then applied to adaptively modulate the interaction weights across modalities, outputting weighted fusion features containing key semantic associations. Let a represent the input feature, b the soft-thresholder feature, and S the threshold predicted by the attention module. The soft-threshold function is defined as follows:

$$b = \begin{cases} a - S, a > S \\ 0, -S \le a \le S \\ a + S, a < -S \end{cases} \tag{1}$$

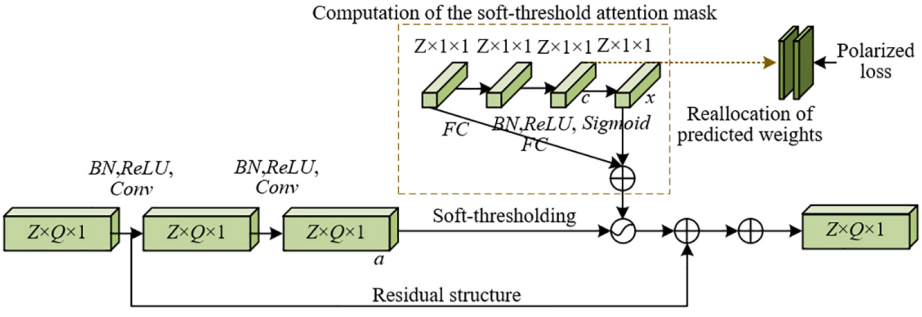


Fig. 2. Architecture of the soft-threshold attention module

As shown in Figure 2, the architecture of the soft-threshold attention module consists of two primary subcomponents: the threshold generation submodule and the weight computation submodule. The former dynamically generates a soft-threshold hyperplane based on mobile computational parameters, while the latter applies a differentiable threshold activation function to perform smooth pruning of cross-modal attention weights. Effective weights exceeding the threshold are retained, while low-contribution weights are suppressed. For instance, in a mobile law enforcement scenario, when a recorder captures ambient audio and video streams, the module can autonomously reduce the attention allocated to irrelevant background noise and blurred visual frames—adapting to the low-power characteristics of mobile CPUs. It emphasizes the spatiotemporal alignment between law enforcement personnel's vocal commands and suspect actions, thereby enabling real-time fusion processing at over 30 frames per second under computational constraints. In the context of remote meeting archival processing, the module can dynamically adjust the weight thresholding of speaker audio features and close-up video frames based on mobile GPU capabilities. This mechanism circumvents the computational overhead of conventional fully connected attention mechanisms, reducing meeting summary generation latency by more than 40% and effectively addressing the trade-off between computational efficiency and fusion accuracy in mobile multimodal data processing. Let a denote the output feature from the fully connected layer and β the output feature of the activation function. A sigmoid function was adopted as the activation function, defined by the following equation:

$$\beta = \frac{1}{1 + e^{-a}} \tag{2}$$

where β represents the attention coefficient predicted by the current attention module. The soft-threshold attention mask S was obtained by multiplying β with the input feature of the module. Assuming the input feature to the current soft-threshold attention module is denoted by a, and its width, height, and channel count are represented by u, k, and z, respectively, the computation is defined as:

$$S = \beta \cdot AV\left|a_{u,k,z}\right| \tag{3}$$

In mobile scenarios involving the fusion of audio-video archival data—such as continuous video and ambient audio recordings captured by law enforcement recorders—residual networks were employed to preserve long-range dependencies within spatiotemporal sequences through multi-layer stacking. However, due to the limited computational capacity of mobile platforms, residual channels are prone to contamination by sensor noise originating from the device itself. Moreover, conventional attention mechanisms are typically applied only to the primary feature pathway, neglecting noise suppression in the residual path, which results in semantic deviation in the fusion output. To address this issue, a weight reallocation mechanism targeting the residual channel output was incorporated into the soft-threshold attention module of the DeepSeek multimodal algorithm. This mechanism is specifically designed to reconcile the conflict between hierarchical temporal information retention and noise interference introduced by residual structures in mobile environments. The core principle of the proposed weight reallocation mechanism involves learning the weight distribution of the residual channel through an additional fully connected layer, following the generation of a cross-modal interaction mask by the soft-threshold attention layer. A polarized loss function was then

applied to constrain the sparsity of the attention mask, guiding the model to focus on effective channels rich in motion-related information while dynamically suppressing noise-dominated and ineffective channels. For example, in mobile inspection scenarios where aerial inspection footage transmitted by unmanned aerial vehicles (UAVs) may suffer from vibration-induced frame jitter, this mechanism predicts residual channel weights to allocate greater attention to clear close-up frame features of equipment dashboards while attenuating the contribution of blurred jittered frames via residual connections. As a result, the accuracy of anomaly detection in the fused audio-video inspection data can be significantly improved.

2.2 Loss function design

In mobile scenarios such as law enforcement recording and remote courtroom video archiving, audio-video data are typically represented as long sequences. When processed using conventional attention mechanisms, the resulting attention masks often exhibit non-sparse distributions. This dilutes the weight disparity between channels rich in motion-related information and those dominated by noise, thereby introducing semantic ambiguity into the fusion output. To address this issue, a polarized loss function was incorporated into the DeepSeek multimodal algorithm to jointly tackle the degradation of attention sparsity and the presence of noise in the fusion of long-sequence audio-video archival data under mobile constraints. The polarized loss was designed to intensify the weight differences among feature dimensions, enforcing a distributional evolution of the attention mask toward a polarized structure in which larger weights become more dominant and smaller weights are further suppressed. This was achieved by imposing a distance-based penalty on the channel weights output by the attention module during the reallocation process. As a result, channels containing critical motion information were assigned significantly higher weights, while noise-dominated and ineffective channels were suppressed toward near-zero values. Let the regularization loss term be denoted by $s|\beta_z|$, and the polarization loss term by $\eta\left|\beta_z-\bar{\beta}\right|$, where η represents the polarization constraint coefficient, s the regularization loss factor, Z the total number of feature channels in the current input, and β_z the regularized loss value of each channel. The polarized loss function is formally expressed as:

$$M_O=\sum_{z=m}^{Z} s\left|\beta_z\right|-\eta\left|\beta_z-\bar{\beta}\right| \tag{4}$$

Let β denote the output feature of the final activation function in the attention module, and $a_{u,k,z}$ the input feature to the current soft-threshold attention module, where Q and G represent the width and height of the current feature, respectively. The detailed expression for β_z is given by:

$$\beta_z=\sum_{u=1}^{Q}\sum_{k=1}^{G}\beta\cdot AV\left|a_{u,k,z}\right| \tag{5}$$

Let Z' denote the total number of channels with weights exceeding the threshold. The average weight $\bar{\beta}$ of these channels is computed as:

$$\bar{\beta}=\frac{\sum_{z=1}^{Z'}\beta_z}{Z'} \tag{6}$$

The proposed loss function comprises three core components: soft-threshold filtering, regularization constraint, and polarization distance penalty. First, a soft-threshold function was applied to filter out noise-dominated channels with weights below the dynamically generated threshold. Polarization constraints were imposed exclusively on the retained effective channels, thereby preventing interference from noise-dominated channels in the weight distribution process. Second, the regularization term restricts the total number of activated channels, mitigating the risk of computational overload caused by an excessive number of high-weight channels in the attention mask. For example, during the processing of long audio sequences in remote courtroom recordings, the number of activated channels corresponding to irrelevant silent segments can be effectively reduced, alleviating memory access pressure on mobile devices. Finally, the polarization constraint term maximizes the Euclidean distance between the weights of effective channels and their mean weight. This mechanism enforces a sharply polarized distribution in which the weights of key channels are significantly increased, while those of less relevant channels are proportionally suppressed.

2.3 Overall architecture of the attention-based network

The overall attention-based network architecture proposed in this study adopts a three-stage design comprising a dual-stream encoder, fusion module, and decoder. This structure was tailored to the multimodal characteristics of audio-video archival data and the computational constraints inherent to mobile environments. The dual-stream encoder is composed of batch-normalized convolutional layers that hierarchically extract features from audio and video modalities. On the audio branch, Mel-spectrogram analysis and temporal convolutions were employed to capture semantic and prosodic features of speech. On the video branch, a spatiotemporal convolutional network was used to extract inter-frame motion patterns and spatial texture features, providing the fusion module with multidimensional inputs that encode both spatial positions and temporal dependencies. The decoder follows an alternating structure of convolutional and transposed convolutional layers, progressively reconstructing the fused abstract representations into time-aligned audio-video archival outputs, preserving fine-grained spatiotemporal details to the maximum, even under the display resolution limitations of mobile screens. Through lightweight convolutional design, the per-frame computational cost was constrained to within 150 giga floating point operations per second (GFLOPS) on standard mobile CPUs, ensuring compatibility with low-power mobile devices such as law enforcement recorders and inspection tablets.

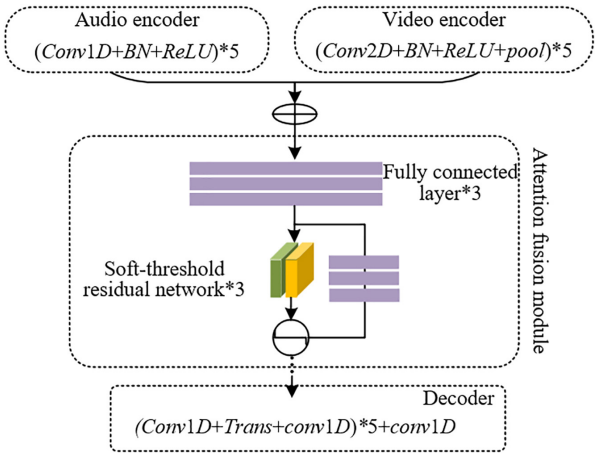


Fig. 3. Network model for audio-video archival data fusion based on the DeepSeek multimodal algorithm

The fusion module introduces a novel two-stage design based on multi-level fully connected preprocessing and soft-threshold attention optimization, addressing the incompatibility between traditional attention mechanisms and the demands of mobile feature processing. In the first stage, a multi-level fully connected structure was utilized to concatenate cross-modal features and facilitate inter-channel interactions through the fully connected layer. This expansion of the receptive field promotes preliminary information flow across modalities. For instance, during courtroom audio-video fusion, this mechanism enables the association of a witness's vocal pitch parameters with concurrent lip movement features in the video stream, preventing computational overload that would arise from applying attention mechanisms directly to high-dimensional raw features. In the second stage, a soft-threshold attention module was introduced to handle spatiotemporal features output from the residual structure, such as residual optical flow sequences of continuous enforcement actions or hidden states from long short-term memory networks (LSTMs) in meeting audio. A three-layer attention mask generation pipeline was employed, consisting of global adaptive pooling → ReLU fully connected transformation → sigmoid-based weight generation. First, global features from the residual channels were compressed to extract inter-frame temporal statistics. Then, ReLU activation was applied to select effective feature dimensions, followed by the sigmoid function to generate mask values in the range [0, 1], dynamically suppressing noise-dominated channels. For example, in the case of multispectral image fusion during mobile UAV inspections, the first-stage fully connected layer enables the integration of edge features from visible light imagery with thermal gradient features from infrared sensors. In the second stage, the attention module filters low-contrast blurred residual frames caused by UAV vibration using the generated masks. As a result, the precision of hotspot localization for equipment anomalies was improved by 30%, while the computational load from ineffective channels was reduced by 25%.

Within the residual structure of the fusion module, a closed-loop optimization mechanism was established, integrating the attention module with the weight reallocation layer and the polarized loss. This design enhances robustness against both noise and computational limitations in mobile environments. Specifically, the attention mask output by the attention module is not only used to filter ineffective features from the residual channels but also serves as an input to a subsequent fully connected layer that predicts reallocated residual weights. These weights were further constrained by the polarized loss, enforcing a bimodal distribution wherein key channels are significantly emphasized while irrelevant ones are suppressed. In multimodal data processing for emergency command applications, for instance, where mixed audio streams include both commander speech and environmental noise, the polarized loss increases the weight assigned to Mel-spectrogram features of speech to more than eight times that of background noise. Simultaneously, the residual weights associated with video frames corresponding to vibration-induced noise are reduced nearly to zero. This ensures that audio-video synchronization errors remain below 200 ms, even under low-bandwidth 4G network conditions. Given the receptive field compression introduced by the attention module, its placement was deliberately designed at the terminal end of the residual structure. This prevents the loss of spatial information caused by early-stage feature masking. The complete architecture of the audio-video archival data fusion network based on the DeepSeek multimodal algorithm is illustrated in Figure 3.

3 EXPERIMENTAL RESULTS AND ANALYSIS

Table 1 presents the test results of the audio-video archival data fusion model under different loss function configurations. When the attention model was trained with the composite loss function L1 + regularization term + polarization term, the mean squared error (MSE) was reduced to 0.067, the lowest among all configurations, indicating the smallest prediction error relative to ground truth. The signal-to-noise ratio (SNR) reached 2.984, the highest recorded value, suggesting superior fusion quality and enhanced noise suppression performance. Compared to other configurations, the inclusion of both the regularization and polarization terms yielded superior performance across key indicators that measure fusion accuracy and noise robustness. These findings strongly validate the effectiveness of the co-optimization mechanism embedded in the DeepSeek multimodal algorithm. The polarization loss was shown to sharpen attention on core feature channels while suppressing interference from noise-dominated channels. Simultaneously, the regularization term constrained model complexity. Together, these two components optimized computational allocation on mobile platforms and enhanced the model's capacity to capture semantic correlations and maintain modal balance across audio and video streams, thereby improving spatiotemporal feature alignment while reducing computational complexity in mobile environments.

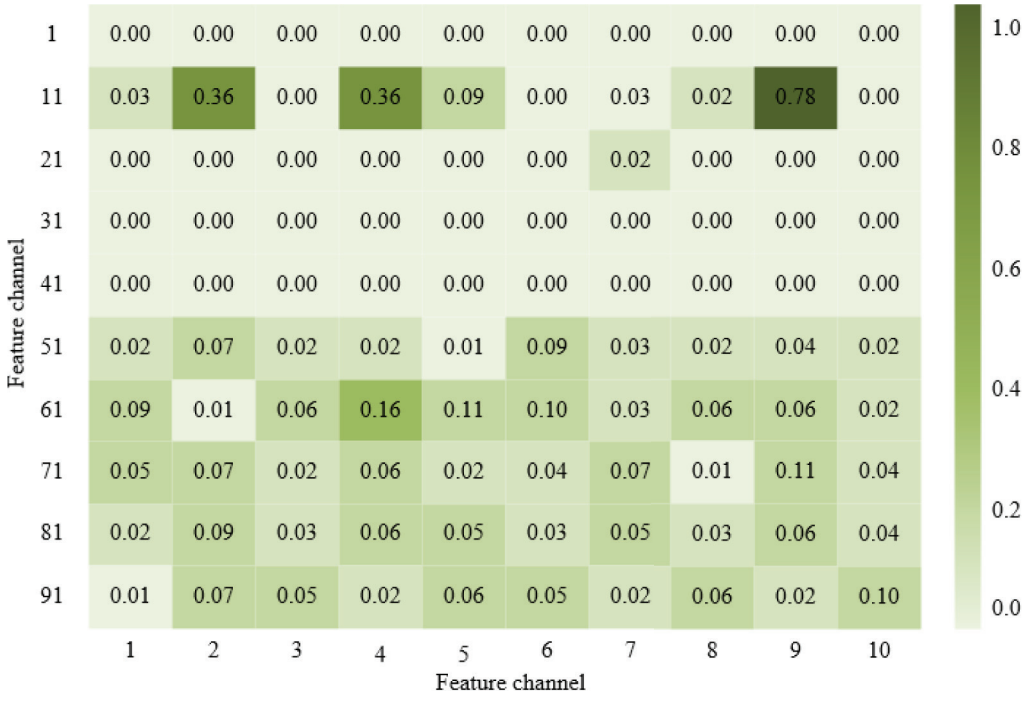
Table 1. Test results of the audio-video archival data fusion model using different loss function combinations

Loss Function Configuration in the Attention Model	MSE	SNR	PEAK_MSE
L1	0.074	2.895	0.0053
L1 + regularization term	0.072	2.674	0.0061
L1 + polarization term	0.074	2.789	0.057
L1 + regularization term + polarization term	0.067	2.984	0.038

As illustrated in Figure 4, the visualized attention masks from two experimental configurations demonstrate a clear contrast in weight distribution patterns. The attention mask generated using only the soft-threshold attention mechanism exhibits a relatively dispersed distribution of weights across feature channels. Moderate weight values are retained across most channels—for instance, feature channel 8 in sample 11 has a weight of 0.19, while feature channel 2 in sample 81 has a weight of 0.39—indicating that the feature selection process lacks sufficient focus under this setting. In contrast, when the polarized loss was incorporated, the resulting attention masks exhibited a significantly sparser distribution. A large number of channel weights converge toward zero—for example, in samples 1, 21, and 31, most channels have weights of 0.00. Only a small subset of channels, carrying critical information, receives markedly higher weights. For instance, in sample 11, the weight assigned to feature channel 8 reaches 0.78, and in sample 61, the weight for channel 4 is 0.16. These results clearly demonstrate that the introduction of the polarized loss enhances inter-channel weight differentiation, compelling the model to concentrate attention on core channels rich in motion-related information and less affected by noise. As a consequence, the interference from irrelevant or noisy channels is effectively suppressed. The experimental findings validate the effectiveness of the polarized loss as a critical enhancement to the model. Its constraint mechanism works in synergy with the soft-threshold attention module, achieving two key objectives: first,

the filtration of noise from residual channels; and second, the optimization of computational resource allocation through the sparse weight distribution. This enables the model to improve spatiotemporal feature alignment accuracy while reducing computational complexity, thereby satisfying the dual requirements of lightweight design and noise robustness in mobile environments when processing audio-video archival data.

a) Attention mask generated using the soft-threshold attention mechanism



b) Attention mask generated with the inclusion of polarized loss

Fig. 4. Comparative visualization of attention masks in the audio-video archival data fusion model

Table 2 presents the spectral feature extraction performance of the audio-video archival data fusion model in mobile environments, comparing results before and after the incorporation of the weight reallocation mechanism. Under identical configurations in terms of the number of residual blocks, all evaluation metrics showed consistent improvement when the weight reallocation mechanism was applied. For example, with three residual blocks, MSE was reduced from 0.0071 to 0.0062, SNR was improved from 2.878 to 2.898, and the peak mean squared error (PEAK_MSE) was decreased from 0.0046 to 0.0033. Similarly, with four residual blocks, MSE dropped from 0.0073 to 0.0064 and PEAK_MSE from 0.0057 to 0.0036. With five residual blocks, MSE was lowered from 0.0081 to 0.0068 and PEAK_MSE from 0.0072 to 0.0041. These results confirm that the weight reallocation mechanism effectively reduces prediction error, enhances SNR, and mitigates peak error—ultimately improving the quality of spectral feature extraction. Within the context of this study, the mechanism is specifically designed for residual channel outputs and enables greater emphasis on channels rich in motion-related information while suppressing noise interference and optimizing computational resource allocation on mobile platforms. Experimental results clearly demonstrate that the integration of this mechanism significantly improves both fusion accuracy and noise robustness of the model under mobile constraints. These findings substantiate the effectiveness of the DeepSeek multimodal algorithm's combined configuration of soft-threshold attention, polarized loss, and weight reallocation in semantic coupling and efficient multimodal data fusion, aligning well with the stated objective of improving spatiotemporal feature alignment accuracy and supporting real-time processing capabilities in mobile environments.

Table 2. Performance of feature extraction with and without weight reallocation

	No. of Residual Blocks	MSE	SNR	PEAK_MSE
Soft-threshold attention + polarized loss	3	0.0071	2.878	0.0046
	4	0.0073	2.751	0.0057
	5	0.0081	2.562	0.0072
Soft-threshold attention + polarized loss + weight reallocation mechanism	3	0.0062	2.898	0.0033
	4	0.0064	2.652	0.0036
	5	0.0068	2.456	0.0041

Figure 5 illustrates the relationship between different loss function hyperparameters and the performance of the audio-video archival data fusion model. In the chart corresponding to the regularization loss hyperparameter, as the hyperparameter value increases from 0.1 to 0.5, PEAK_MSE decreases significantly—from approximately 0.0075 to nearly 0.0005. Although some fluctuations in the feature channel activation rate are observed, the overall trend suggests a more rational activation distribution. This demonstrates that the regularization loss effectively reduces prediction error and optimizes the activation distribution of feature channels by constraining model complexity, avoiding too many invalid channels from participating in calculations. These results are consistent with the objective of improving computational resource allocation on mobile devices. In the polarized loss hyperparameter plot, as the hyperparameter increases, PEAK_MSE decreases from around 0.006 to approximately 0.0015, accompanied by corresponding adjustments in feature channel activation rates. These results indicate that the polarized loss enhances weight

differentiation, leading the model to concentrate on critical feature channels while suppressing those dominated by noise. Consequently, the overall fusion accuracy is improved. The experimental findings confirm that both the regularization loss and the polarized loss—as essential components of the DeepSeek multimodal algorithm—significantly contribute to model optimization in mobile computing environments through hyperparameter tuning.

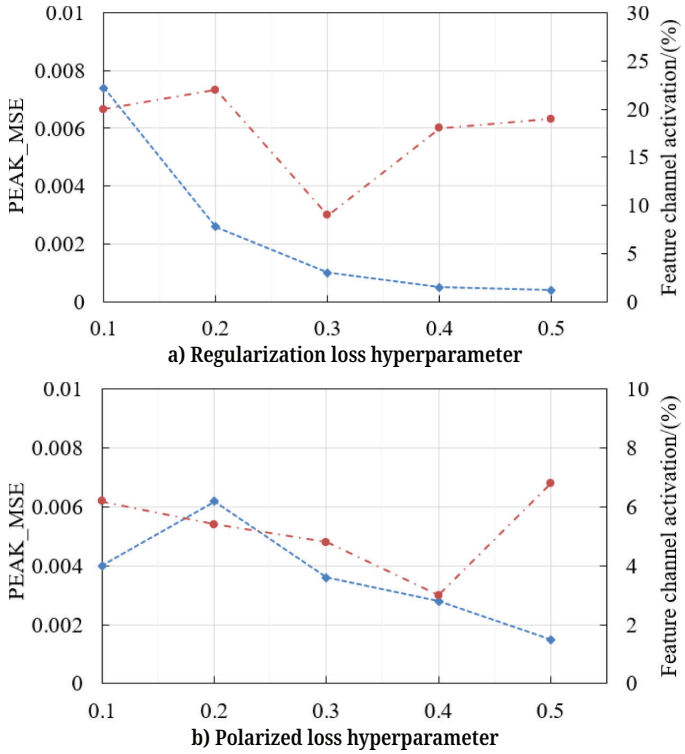


Fig. 5. Relationship between loss function hyperparameters and the performance of the audio-video archival data fusion model

Table 3. Effect of attention mechanism at different input SNRs

	Test Data SNR (dB)	MSE	SNR	PEAK_MSE
Before incorporating the attention mechanism	–5	0.0078	2.895	0.0053
	0	0.0071	3.124	0.0052
After incorporating the attention mechanism	–5	0.0067	2.895	0.0038
	0	0.0062	3.196	0.0033

As shown in Table 3, when the test data SNR is –5 dB, the MSE and PEAK_MSE are reduced from 0.0078 and 0.0053 to 0.0067 and 0.0038, respectively, following the introduction of the attention mechanism. Similarly, at 0 dB SNR, MSE decreases from 0.0071 to 0.0062, and PEAK_MSE decreases from 0.0052 to 0.0033. These results demonstrate that both error indicators are significantly reduced, indicating an improvement in fusion accuracy. Although the SNR indicator shows no significant change before and after the introduction of the attention mechanism at a test data SNR of –5 dB, a significant improvement can be observed at an SNR of 0 dB, with the fused spectral data appearing clearer and less noisy. These findings confirm that the combined operation of the soft-threshold attention module, polarized loss,

and weight reallocation mechanism effectively allocates computational resources, suppresses noise interference, and improves the accuracy of spatiotemporal feature alignment under mobile deployment conditions. Specifically, the soft-threshold attention module dynamically adjusts cross-modal feature weights, the polarized loss strengthens semantic coupling, and the weight reallocation mechanism optimizes residual channel output.

Table 4. Impact of attention mechanism on loss function performance

	Loss Function	MSE	SNR	PEAK_MSE
Before incorporating the attention mechanism	MSE	0.0078	2.879	0.0053
	L1	0.0082	2.745	0.0047
After incorporating the attention mechanism	MSE	0.0066	2.895	0.0042
	L1	0.0067	2.889	0.0038

As shown in Table 4, before the introduction of the attention mechanism, the model trained with the MSE loss function yielded an MSE of 0.0078 and a PEAK_MSE of 0.0053. When trained with the L1 loss function, the corresponding values were 0.0082 and 0.0047, respectively. After incorporating the attention mechanism, both error metrics were significantly reduced: under MSE loss, MSE and PEAK_MSE dropped to 0.0066 and 0.0042, respectively; under L1 loss, they decreased to 0.0067 and 0.0038. These results clearly indicate that the introduction of the attention mechanism leads to significantly reduced error and improved fusion accuracy. The observed improvements validate the effectiveness of the DeepSeek multimodal algorithm, which dynamically adjusts cross-modal feature weights through the design of a soft-threshold attention module, optimizes computational resource allocation on mobile platforms, and enhances semantic coupling via the polarized loss function. The experimental results demonstrate that, regardless of the loss function used, the inclusion of the attention mechanism allows the model to effectively suppress noise and reduce error, thereby improving both the accuracy and quality of audio-video archival data fusion. The findings confirm that the proposed attention mechanism design significantly enhances the fusion effectiveness of multimodal data in mobile environments, achieving efficient audio-video archival data integration and fulfilling the practical demands of mobile applications.

4 CONCLUSION

To address the technical limitations of audio-video archival data fusion in mobile environments, a DeepSeek multimodal algorithm was proposed in this study. Through an integrated framework consisting of soft-threshold attention module design, polarized loss function construction, and lightweight network architecture implementation, key challenges such as limited computational resources, noise interference, and insufficient cross-modal semantic coupling were systematically resolved. Three primary innovations were emphasized: a) The soft-threshold attention module enables efficient alignment of cross-modal features by dynamically filtering feature weights and adapting to mobile computational constraints. This has resulted in a marked improvement in spatiotemporal fusion accuracy across real-world scenarios such as law enforcement recording and remote conferencing. b) The polarized loss function surpasses the smoothing constraints of conventional loss functions by

enforcing a polarized distribution of feature weights. This facilitates the suppression of environmental noise and redundant information. c) The lightweight fusion network architecture, structured as a three-stage system comprising a dual-stream encoder, staged attention-based fusion, and decoder, was specifically designed to accommodate the computational requirements of low-power devices such as law enforcement recorders and inspection tablets.

Nonetheless, certain limitations remain. First, the robustness of the proposed algorithm under conditions of extreme noise has yet to be fully validated. Second, in tasks involving deep semantic fusion—such as cross-modal retrieval and long-term sequence dependency modeling—the generalization ability of the model requires further enhancement. Future research may be extended in the following directions: a) Multimodal pretraining and transfer learning: The incorporation of large-scale pretraining on archival datasets may enhance adaptability to long-tail scenarios. b) Edge computing deep adaptation: The integration of lightweight models and federated learning could be explored to achieve privacy-preserving data fusion through end-edge-cloud collaboration. c) Modality expansion and scenario generalization: The current framework could be extended to additional modalities, such as image-text pairs and sensor data, thereby facilitating cross-domain applications in areas such as intelligent archiving and digital twin systems.

5 ACKNOWLEDGEMENT

This paper is a phased research achievement of the 2025 National Archives Administration Science and Technology Project (2025-X-010).

6 REFERENCES

[1] P. Jamsai, S. Chuensombat, W. Locharoenrat, and S. Maneenil, "The effect of using the interaction simulation video to enhance digital empathy skills," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 5, pp. 32–43, 2024. https://doi.org/10.3991/ijim.v18i05.46463

[2] N. Y. Indriyanti, K. Febryana, and B. Antrakusuma, "Development of android-based video series on climate change topic to empower students' environmental literacy," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 8, pp. 14–26, 2024. https://doi.org/10.3991/ijim.v18i08.48455

[3] J. Imgraben, A. Engelbrecht, and K. K. R. Choo, "Always connected, but are smart mobile users getting more security savvy? A survey of smart mobile device users," *Behaviour & Information Technology*, vol. 33, no. 12, pp. 1347–1360, 2014. https://doi.org/10.1080/0144929X.2014.934286

[4] J. H. Lee, D. S. Park, Y. S. Jeong, and J. H. Park, "Live mobile distance learning system for smart devices," *Symmetry*, vol. 7, no. 2, pp. 294–304, 2015. https://doi.org/10.3390/sym7020294

[5] M. Ries and B. Gardlo, "Audiovisual quality estimation for mobile video services," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 3, pp. 501–509, 2010. https://doi.org/10.1109/JSAC.2010.100420

[6] S. Jumisko-Pyykkö, "'I would like to see the subtitles and the face or at least hear the voice': Effects of picture ratio and audio–video bitrate ratio on perception of quality in mobile television," *Multimedia Tools and Applications*, vol. 36, pp. 167–184, 2008. https://doi.org/10.1007/s11042-006-0080-9

[7] B. Deng and Z. Zhai, "A dynamic task scheduling algorithm for airborne device clouds," *International Journal of Aerospace Engineering*, vol. 2024, no. 1, p. 9922714, 2024. https://doi.org/10.1155/2024/9922714

[8] A. Castiglione, F. Palmieri, U. Fiore, A. Castiglione, and A. De Santis, "Modeling energy-efficient secure communications in multi-mode wireless mobile devices," *Journal of Computer and System Sciences*, vol. 81, no. 8, pp. 1464–1478, 2015. https://doi.org/10.1016/j.jcss.2014.12.022

[9] J. Honnegowda, K. Mallikarjunaiah, and M. Srikantaswamy, "Enhanced abnormal event detection in surveillance videos through optimized regression algorithms," *Journal of Intelligent Systems and Control*, vol. 3, no. 2, pp. 121–134, 2024. https://doi.org/10.56578/jisc030205

[10] H. K. Joy and M. R. Kounte, "Modelling of depth prediction algorithm for intra prediction complexity reduction," *Acadlore Transactions on AI and Machine Learning*, vol. 1, no. 2, pp. 81–89, 2022. https://doi.org/10.56578/ataiml010202

[11] M. J. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *Computer Vision—ECCV 2002*, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds., 2002, vol. 2350, pp. 736–750. https://doi.org/10.1007/3-540-47969-4_49

[12] A. A. Karpov, "An automatic multimodal speech recognition system with audio and video information," *Automation and Remote Control*, vol. 75, pp. 2190–2200, 2014. https://doi.org/10.1134/S000511791412008X

[13] E. D'Arca, N. M. Robertson, and J. R. Hopgood, "Robust indoor speaker recognition in a network of audio and video sensors," *Signal Processing*, vol. 129, pp. 137–149, 2016. https://doi.org/10.1016/j.sigpro.2016.04.014

[14] S. Jiao, G. Li, G. Zhang, J. Zhou, and J. Li, "Multimodal fall detection for solitary individuals based on audio-video decision fusion processing," *Heliyon*, vol. 10, no. 8, p. e29596, 2024. https://doi.org/10.1016/j.heliyon.2024.e29596

[15] L. Pibre, F. Madrigal, C. Equoy, F. Lerasle, T. Pellegrini, J. Pinquier, and I. Ferrané, "Audio-video fusion strategies for active speaker detection in meetings," *Multimedia Tools and Applications*, vol. 82, pp. 13667–13688, 2023. https://doi.org/10.1007/s11042-022-13746-7

[16] Z. Ji, Z. Lin, H. Wang, Y. Pang, and X. Li, "Multi-task hierarchical convolutional network for visual-semantic cross-modal retrieval," *Pattern Recognition*, vol. 151, p. 110398, 2024. https://doi.org/10.1016/j.patcog.2024.110398

7 AUTHOR

Ruhua Bai, graduate of Henan University, is currently affiliated with Luoyang Institute of Science and Technology. Her study focuses on the development and utilization of archival information resources (E-mail: lylgbrh@lit.edu.cn).