Kindergarten Safety Education Recommendation Based on Multimodal Fusion and ALBERT Lightweight Model

Ping Zhang[1], Xiu Fan[2]
[1]School of Humanities and Management, Xi'an Traffic Engineering Institute, Xi'an 710300, Shaanxi, China
[2]Shaanxi Radio & Television Station, Xi'an 710063, Shaanxi, China
Corresponding author: Xiu Fan (e-mail: aicode123@163.com)

ABSTRACT Current mainstream multimodal recommendation models for kindergarten safety education usually have large parameter scales and high computational overhead, which makes them difficult to deploy on resource-constrained edge devices. Similar deployment constraints also exist in engineering systems involving wireless sensing, electromagnetic environment monitoring, and antenna-assisted edge perception, where real-time inference and low resource consumption are required. To balance lightweight design and multimodal information integration, this paper proposes a kindergarten safety education recommendation framework based on multimodal fusion and an ALBERT lightweight model. The framework uses ALBERT-base-v2 with frozen embedding layers and fine-tunes only the last two Transformer encoder layers as the text encoder. A lightweight ResNet-18 and WaveNet are adopted for image and audio feature extraction, respectively. Trimodal semantic alignment is achieved through 128-dimensional linear projection, and a low-overhead gating mechanism is introduced to dynamically integrate modal features. Finally, a dual-tower architecture and Bayesian Personalized Ranking (BPR) loss are used to realize personalized recommendation. Experimental results show that the proposed model achieves an inference latency of 86.67 ms, a model size of 87.3 MB, an average HR@5 of 91.4%, and an average age-appropriateness score of 4.32. The results indicate that the model effectively balances deployment efficiency, multimodal representation capability, and adaptability to children's development, providing a practical solution for intelligent kindergarten safety education systems and lightweight edge recommendation applications related to electromagnetic-aware sensing environments.

INDEX TERMS albert lightweight model, multimodal fusion, educational recommendation, edge intelligence, electromagnetic sensing

I. INTRODUCTION

THE kindergarten stage is a critical period for children to develop safety awareness, and personalized and contextualized safety education content is urgently needed [1-3]. In recent years, multimodal learning has shown potential in the field of educational recommendations due to its ability to integrate multiple sources of information such as text, images, and voice, paralleling its growing application in the textile industry for comprehensive material analysis and defect identification based on visual and spectral data [4-6]. However, existing research generally relies on multimodal pre-training models with large parameter counts and high computational overhead, making it difficult to deploy on kindergarten edge devices with limited computing resources and network conditions [7,8]. At the same time, although lightweight language models such as ALBERT have the advantage of low-resource adaptation, they only process text modalities and cannot effectively utilize key non-text information such as picture book illustrations, safety sign images, or voice prompts. The recommended content is out of touch with children's cognitive level and perceptual habits, limiting the effectiveness of safety education [9].

This paper focuses on recommended scenarios for kindergarten safety education for preschoolers aged 3–6 years. The research subjects include three core types of multimodal content: safety education text, accompanying images, and safety audio. This content is highly age-sensitive and modally complementary—young children primarily rely on visual and auditory input to understand abstract safety concepts, and text alone is insufficient to meet their cognitive needs [10,11]. Many scholars have verified the importance of this object from the perspective of the intersection of developmental psychology and educational technology: Zashchirinskaia O. V. used eye movement experiments to understand the visual image sensitivity of preschool children [12]; safety education texts such as nursery rhymes play a key role in the learning ability and learning efficiency of preschool children. Nursery rhymes can be used as a

medium to assist in memorizing knowledge points [13-15]; the coordinated presentation of images and voices (such as animation) can enable children to improve their mastery of knowledge and skills when learning [16]; multimodal learning has a positive effect on preschool children and helps to improve children's interest in learning [17,18]. Therefore, a lightweight recommendation system capable of integrating multimodal semantics and meeting edge deployment constraints is of practical significance for improving age-appropriate matching and accessibility. Regarding multimodality, existing work has attempted to combine various models to support students' multimodal learning, or to use a dual-stream architecture to achieve cross-modal interaction [19,20]. Ren X et al. proposed a learning recommendation based on Long Short-Term Memory (LSTM) and attention mechanism, but lacked performance in the short text scenario of young children [21]. Ryu M. et al. and Cho I. et al. performed knowledge distillation on Bidirectional Encoder Representation from Transformers (BERT), but did not optimize and adapt it for resource-poor edge devices [22,23]. Safitri S. et al. used audio as a medium to activate children's memory, but did not align it with the visual modality [24]. Uppada S. K. et al. proposed a multimodal model based on images and text, but ignored the role of speech [25]. Existing lightweight solutions generally do not consider the joint constraints of end-to-end deployment volume, inference latency, and age-appropriate matching, making it difficult to implement in real kindergarten environments. This paper aims to construct a lightweight multimodal safety education recommendation framework for resource-constrained kindergartens. Based on ALBERT-base-v2, an efficient text encoder is constructed. A lightweight ResNet-18 and WaveNet are applied to process images and audio, respectively, and trimodal semantic alignment is achieved through a 128-dimensional shared semantic space. A gated fusion mechanism is further designed to dynamically weight the contributions of each modality. Finally, a dual-tower architecture and BPR loss are used to achieve efficient personalized recommendations. Experiments demonstrate that this framework outperforms existing lightweight baselines, effectively balancing lightweight design, multimodal fusion efficiency, and child-appropriateness, achieving an optimization balance similar to that required for effective, high-speed defect detection systems in textile manufacturing.

II. METHODS

A. MULTIMODAL INPUT PREPROCESSING

The text input is kindergarten safety education related corpus (such as picture book sentences, children's song lyrics), and the ALBERT-base-v2 tokenizer provided by the Hugging Face Transformers library is used for word segmentation to generate word unit sequences [26]. The sequence length is uniformly truncated or padded to the maximum length $L = 64$, and the output is an integer tensor $x_{\text{text}} \in \mathbb{Z}^L$. At the same time, the corresponding attention mask $m_{\text{text}} \in \{0,1\}^L$ is generated, where $m_i = 1$ is the i-th position equals 1 if it corresponds to a valid token, and 0 otherwise.

Image input, including security logos and animation keyframes, is read in Blue, Green, Red (BGR) format using Open Source Computer Vision Library (OpenCV), converted to Red, Green, Blue (RGB), and uniformly scaled to 224×224 pixels. Channel normalization is then performed:

$$X_{\text{img}}^{(c)} = \frac{X_{\text{raw}}^{(c)}/255.0 - \mu_c}{\sigma_c}, \quad c \in \{R, G, B\} \qquad (1)$$

In Formula 1, $X_{\text{raw}}^{(c)}$ represents the original pixel value, and μ_c and σ_c are the channel mean and standard deviation, respectively, to ensure consistency with the input distribution of the pre-trained visual model. The audio input consists of a 3-second safety prompt voice message. It is first resampled to 16 kHz and converted to mono using the Librosa library. A mel-frequency spectrogram is extracted using a Hann window (25 ms frame length, 10 ms hop size) is used, and a 512-point Fast Fourier Transform (FFT) to obtain a 64-channel mel-frequency filter bank response [27-29]. The time series length of the spectrogram is fixed $T = 300$ (corresponding to 3 seconds).Anything less than 3 seconds is padded with zeros, and anything exceeding 3 seconds is truncated, resulting in a tensor $X_{\text{audio}} \in \mathbb{R}^{64\times T}$.

B. LIGHTWEIGHT TEXT ENCODER CONSTRUCTION

The text encoder is based on the ALBERT-base-v2 architecture provided by the Hugging Face transformers library [30,31]. To reduce computational overhead and adapt to edge deployment, a layered parameter freezing strategy is implemented: the embedding layers (including token, position, and segment embeddings) are fixed, and only the last two Transformer blocks and the pooler layer are fine-tuned. The term "fine-tuning only the top layers" (used in this paper) can be interpreted as "fixing all embedding layers of ALBERT-base-v2 and only updating gradient on the 11th and 12th Transformer encoder modules plus pooling layer", where ALBERT-base-v2 has 12 encoder layers. Therefore, when we refer to the "top layers", we are referring to 11 & 12 as opposed to 12 (classification output layer) in the definition. The input text is processed by the ALBERT tokenizer to generate input_ids and attention_mask. After input into the model, it is passed through the first 10 frozen Transformer blocks, with only the 11th and 12th layers participating in gradient updates. Finally, the hidden state $h_{\text{cls}} \in \mathbb{R}^{768}$ corresponding to the [CLS] token is used as the sentence-level semantic representation.

To control the scale of training parameters, the embedding layer is completely frozen to avoid overfitting on the small-scale safety education corpus. At the same time, the trainable layers are limited to reduce memory usage and backpropagation computation. The frozen embeddings will allow us to use the different trained layers to create a model that has a small number of trainable parameters and only requires a limited amount of computational resources when

performing inference, thereby enabling us to produce the most efficient model for this application. The freezing of the model's embedding layers will also reduce the amount of memory consumed by the parameters of the model (the stored values), because the parameters for the embedding layers will remain constant at their pre-trained values during inference. Not having to calculate gradients for the parameters of the embedding layers during backpropagation will allow for reduced memory consumption and computational overhead on edge devices during training and inference processes. Additionally, only the last two layers of the model were trainable for this project, further reducing the resources required on edge devices and allowing for the size of the final exported inference model to be controlled. Additionally, the model will be lightweight because the combination of frozen embedding layers, top-layer fine-tuning, low-dimensional alignment, and gated fusion mechanisms all contribute to making the model lightweight. The AdamW optimizer is used during the fine-tuning phase. The sentence vector output is formalized as follows:

$$h_t = \text{Pooler}\left(\text{Transformer}_{11:12}\left(\text{FrozenEmb}(x_{\text{text}}, m_{\text{text}})\right)\right) \tag{2}$$

In Formula 2, $x_{\text{text}} \in \mathbb{Z}^{64}$ is the input ID sequence after word segmentation; $m_{\text{text}} \in \{0,1\}^{64}$ is the attention mask; $\text{FrozenEmb}(\cdot)$ represents the frozen embedding layer mapping; the output dimension is $\mathbb{R}^{64\times768}$; $\text{Transformer}_{11:12}(\cdot)$ represents the 11th to 12th layer trainable Transformer block; $\text{Pooler}(\cdot)$ is a linear transformation followed by Tanh activation; the 768-dimensional vector h_t of the [CLS] position is extracted.

C. CROSS-MODAL FEATURE ALIGNMENT MODULE

To ensure comparability of text, image, and audio features in a unified semantic space, this module performs dimension-wise projection on the raw encoded output of each modality, ultimately mapping it to a 128-dimensional shared embedding space. Text features $h_t \in \mathbb{R}^{768}$ are derived from a lightweight ALBERT encoder and projected through a learnable linear layer $W_t \in \mathbb{R}^{128\times768}$ to produce an alignment vector $h'_t = W_t h_t \in \mathbb{R}^{128}$.

After preprocessing, the image input is fed into a ResNet-18 pre-trained on ImageNet. The last fully connected layer is removed, and the 512-dimensional features after global average pooling are extracted. The features $h_v \in \mathbb{R}^{512}$ are then mapped to $h'_v = W_v h_v \in \mathbb{R}^{128}$ through a linear transformation $W_v \in \mathbb{R}^{128\times512}$. This strategy reuses mature visual representations, avoids training from scratch, and reduces data and computing power requirements [32-34]. ResNet-18 was chosen as the visual encoder due to its ability to provide a good balance between performance and being lightweight. ResNet-18 has been pre-trained on ImageNet, so it produces a superior feature representation compared with lighter encoders like MobileNet, this is beneficial for recognition tasks involving complex safety education imagery containing many different symbols/scenes. This was taken into consideration for ResNet-18's architecture, which has been highly-optimized for inference efficiency on mobile devices. Finally, compressing the feature dimensions to 128-dimensional linear and then projecting them through 128-dimensional linear projection helps manage computational/storage overhead by decreasing feature dimension size, so this provides a reliable and efficient means of capturing visual features using ResNet-18.

The audio mel-spectrogram $X_{\text{audio}} \in \mathbb{R}^{64\times301}$ is fed into a lightweight WaveNet architecture: it consists of three layers of causal convolution (kernel size=3, dilation=1, 2, 4), with the number of channels increasing from 64 to 128 and remaining at 128 in the final layer, and each layer is followed by a ReLU activation. The output time series features $Z \in \mathbb{R}^{128\times T}$ are globally average-pooled to obtain $h_a = \frac{1}{T}\sum_{t=1}^{T} z_t \in \mathbb{R}^{128}$, that is, the aligned audio vector h'_a. A simplified WaveNet architecture was adopted instead of MFCC-based or shallow CNN approaches to balance temporal modeling capability and computational cost. By using causal convolutions with a dilated structure, we can model temporal dependencies within the audio spectrogram via time-based correlation which is essential in understanding sequential instruction and emotional intonation in safety education voice prompts. The architecture we designed included three convolutional layers, with very few channels in each layer, and global average pooled layers to extract fixed dimensional representations. The computational cost of the architecture is strictly managed and controlled. In comparison to the fixed features of MFCC, our learnable encoder will extract more and better discriminative features from the limited data, which aligns with the semantics of both the text and the image, all the while maintaining low computational overhead to satisfy real-time requirements for deployment at the edge.

The trimodal alignment process is uniformly expressed as:

$$h'_m = \begin{cases} W_t h_t, & m = \text{text}, \\ W_v h_v, & m = \text{vision}, \\ \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} \text{ReLU}\big(\text{Conv}_3^{(3)} \circ \text{Conv}_2^{(2)} & \\ \qquad \circ\, \text{Conv}_1^{(1)}(X_{\text{audio}})\big)_t & m = \text{audio}. \end{cases} \tag{3}$$

In Formula 3, $h'_m \in \mathbb{R}^{128}$ is the alignment feature of modality m; $\text{Conv}_i^{(d)}$ represents the i-th layer causal convolution dilation rate d; $T = 301$ is the number of spectrogram time steps. This design preserves the discriminative nature of each modality while achieving low-dimensional semantic alignment, providing structurally consistent input for subsequent gated fusion, significantly reducing model complexity, and adapting to edge device memory constraints.

D. DYNAMIC GATING FUSION MODULE

To dynamically integrate the complementary information of text, image, and audio modalities, a weighted fusion strategy based on a gating mechanism is designed. First, the original high-dimensional features of each modality are concatenated: text features $h_t \in \mathbb{R}^{768}$ (from the ALBERT pooler output), image features $h_v \in \mathbb{R}^{512}$ (ResNet-18 global pooling output), and audio features $h_a \in \mathbb{R}^{128}$ (WaveNet pooling vector) are concatenated into $h_{\text{concat}} = [h_t; h_v; h_a] \in \mathbb{R}^{1408}$. This concatenated vector is fed into a two-layer MLP (with a 256-dimensional hidden layer and ReLU activations), which output three-dimensional weighted logits $z = [z_t, z_v, z_a]^\top \in \mathbb{R}^3$.

Subsequently, the Softmax function is applied to z to generate the normalized gating weights:

$$\alpha_m = \frac{\exp(z_m)}{\sum_{m' \in \{t,v,a\}} \exp(z_{m'})}, \quad m \in \{t, v, a\} \tag{4}$$

α_t, α_v, α_a are the dynamic weights of text, vision, and audio modalities, respectively, satisfying $\alpha_t + \alpha_v + \alpha_a = 1$ and $\alpha_m \geq 0$.

The final fusion vector is obtained by weighted summation of the aligned 128-dimensional features:

$$h_{\text{fused}} = \alpha_t h'_t + \alpha_v h'_v + \alpha_a h'_a \tag{5}$$

In Formula 5, $h'_t = W_t h_t \in \mathbb{R}^{128}$, $h'_v = W_v h_v \in \mathbb{R}^{128}$, and $h'_a = h_a \in \mathbb{R}^{128}$ are alignment features. This mechanism enables the model to adaptively adjust modal contributions based on the input content—for example, when the input is pure speech prompts, α_a automatically enhances to avoid interference from invalid modalities.

E. LIGHTWEIGHT RECOMMENDED HEAD DESIGN

The recommendation head adopts a dual-tower structure, modeling user preferences and content representation separately to reduce the complexity of online inference. The user tower input consists of two parts:

(1) a one-hot encoding $a \in \{0, 1\}^3$ of the child's age (corresponding to the three age groups of 3–4 years old, 4–5 years old, and 5–6 years old); (2) the average fused features $\bar{h}_{\text{fused}} = \frac{1}{5} \sum_{i=1}^{5} h_{\text{fused}}^{(i)} \in \mathbb{R}^{128}$ of the user's five most recent interactions. The two are concatenated into $[a; \bar{h}_{\text{fused}}] \in \mathbb{R}^{131}$ and mapped into a user vector $u \in \mathbb{R}^{128}$ through a 128-dimensional single-layer MLP (with ReLU activation).

The content tower input is the multimodal fusion features $h_{\text{fused}} \in \mathbb{R}^{128}$ of the current safety education content, and outputs the content vector $c \in \mathbb{R}^{128}$ through a 128-dimensional Multi-Layer Perceptron (MLP) with the same structure (shared weights are optional and are independent here to preserve modality specificity). The recommendation score is calculated as:

$$s(u, c) = u^\top c \tag{6}$$

The model is optimized using BPR loss [35]. For each user u and its positive i and negative samples j, the loss function is defined as:

$$L_{\text{BPR}} = -\sum_{u \in U} \sum_{i \in I_u^+} \sum_{j \in I \setminus I_u^+} \ln \sigma \left(s(u, c_i) - s(u, c_j) \right) \tag{7}$$

In Formula 7, U is the user set; I_u^+ is the set of content that the user u has positive interactions; I is the entire content library; $\sigma(\cdot)$ is the sigmoid function; c_i and c_j are the content vectors for positive and negative samples, respectively. This loss forces the model to learn relative ranking rather than absolute ratings, which aligns with the implicit feedback nature of recommendation scenarios. The BPR loss curve during the training of the quantized recommendation head is shown in Figure 1.

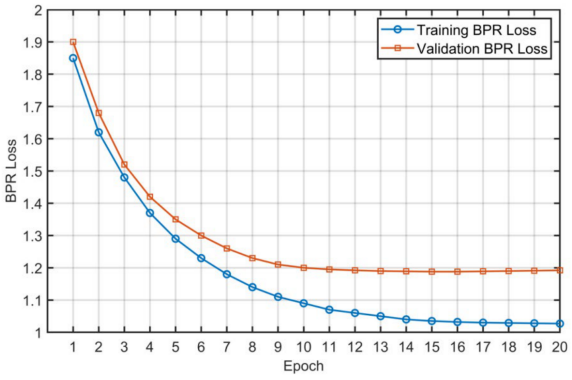

FIGURE 1. BPR loss curve during lightweight recommendation head training

Figure 1 illustrates the BPR loss curve, which exhibits a rapid decline followed by a plateau without obvious oscillation or overfitting, indicating effective learning of implicit preference relationships. The steady decline of the BPR loss indicates effective learning of relative preference ranking.

III. EXPERIMENTS AND VERIFICATION

A. EXPERIMENTAL DESIGN

The experimental data is collected at a partner kindergarten under the guidance of educational experts. It contains 12,000 multimodal samples, each of which includes a safety education text description, a corresponding image, a 3-second audio clip, and the child's age label. To prevent the leakage of user behavior information, the data is stratified by user ID, with a training set, validation set, and test set ratio of 7:1:2. This study was approved by the Ethics Committee of Xi'an Traffic Engineering Institute, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

The baseline models include five categories: (1) Text-only ALBERT, which only uses text modality; (2) BERT + ResNet, which uses a combination of standard BERT and

ResNet-50; (3) short for Vision-and-Language BERT (ViLBERT), a typical two-stream multimodal pre-training model; (4) Learning Cross-Modality Encoder Representations from Transformers (LXMERT), which supports three modalities but has a large number of parameters; (5) MobileViT + DistilBERT, the current lightweight multimodal representative.

The method in this paper: based on ALBERT-base-v2, ResNet-18, and WaveNet, text, image, and audio are encoded, respectively; cross-modal alignment is achieved through 128-dimensional linear projection; a gated fusion mechanism is designed to dynamically weight the three-modal features. Finally, a dual-tower structure and BPR loss are used to achieve personalized recommendations.

Model deployment and inference testing are conducted on an NVIDIA Jetson Nano embedded platform (4 GB Random Access Memory (RAM) and a 128-core Maxwell Graphics Processing Unit (GPU)). The operating system is Ubuntu 18.04, and the inference framework is PyTorch 1.10 + TorchScript. All model inputs undergo a preprocessing process to ensure fair comparison. The AdamW optimizer is used during training. This design demonstrates the practicality and deployment feasibility of the proposed framework in resource-constrained scenarios.

B. RECOMMENDATION HIT RATE

The model's hit rate at 5 (HR@5) is used as a core metric for quantitative evaluation, measuring the model's ability to capture the user's real interactions in the top-5 recommendations. The specific implementation process is as follows: on the test set, for each user, the model generates recommendation scores for all items in the candidate content library based on their historical behavior (their five most recent clicks) and age label. The scores are sorted in descending order, and the top five items are selected to form the recommendation list. If the user's real interaction item in the test set (i.e., the next clicked item) appears in the recommendation list, it is considered a hit; otherwise, it is considered a miss. In the experiments, the HR@5 score is taken as the arithmetic average to avoid excessive dominance of high-frequency users in overall performance and more fairly reflect the model's generalization ability across different children. Each experiment is repeated five times. A comparison of the HR@5 scores of the models across the five rounds of testing is shown in Figure 2.

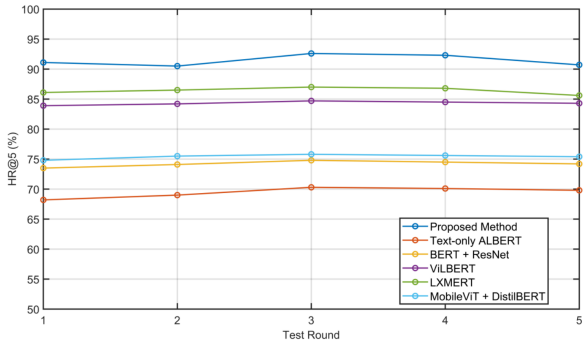

FIGURE 2. HR@5 comparison of each model in 5 rounds of testing

Figure 2 compares HR@5 results across different models. The proposed method maintains HR@5 above 90% across five rounds, with an average of 91.4%. This performance is higher than that of text-only ALBERT (approximately 69.5%), MobileViT + DistilBERT (approximately 75.4%), ViLBERT (approximately 84.3%), LXMERT (approximately 86.4%), and BERT + ResNet (approximately 74.2%). The results indicate that the proposed method improves recommendation accuracy while maintaining lightweight design. This performance advantage stems from the model's effective integration of multimodal information and the coordinated optimization of its lightweight design within the underlying architecture.

C. MODEL INFERENCE LATENCY

The inference latency test process covers the entire chain from raw multimodal input to final recommendation score output, including: text segmentation and encoding, image preprocessing and feature extraction, audio mel-spectrogram generation and encoding, cross-modal alignment, gated fusion, and recommendation head calculation. During testing, 1,000 samples are randomly sampled from the test set, each containing a complete text, image, and audio triplet. For each sample, the time interval from data loading completion to recommendation score output is recorded, using a high-precision system timer to ensure accuracy. To eliminate cold start and cache effects, 200 warm-up inferences are performed before the actual test. During the actual test, system background tasks, GPU dynamic frequency scaling, and swap partitions are disabled to ensure hardware stability. All operations are performed in single-threaded mode, simulating a typical edge device deployment environment with no concurrent resources. The inference latency of each model is shown in Figure 3.

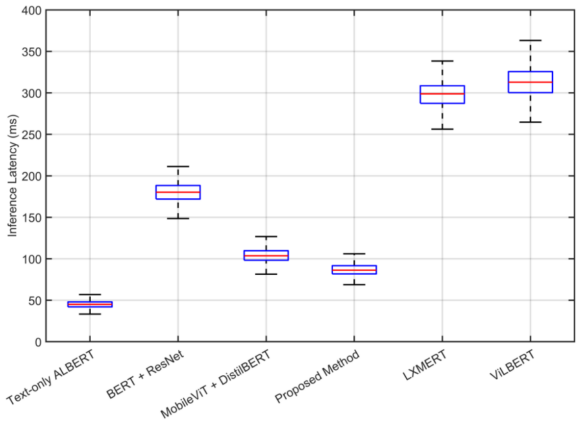

FIGURE 3. Inference latency of each model

Figure 3 shows that the proposed lightweight multimodal model achieves an average latency of 86.67 ms and a minimum latency of 68.96 ms, significantly lower than other multimodal models such as ViLBERT (312.65 ms)

and LXMERT (298.14 ms), and also outperforms the current lightweight representative, MobileViT + DistilBERT (103.96 ms). In comparison, Text-only ALBERT, which processes only text, has the lowest latency (average of approximately 45.02 ms), while BERT + ResNet achieves a latency of 180.4 ms. While maintaining multimodal capabilities, the proposed method achieves efficient inference, validating its real-time advantages in edge device deployment. This reduction in inference latency stems from the proposed method's multiple lightweight strategies in both architectural design and engineering implementation. The text encoder freezes the ALBERT embedding layers and fine-tunes only the top layers, significantly reducing trainable parameters and computational overhead.

D. VERIFICATION OF MULTIMODAL COMPLEMENTARITY

To quantify the actual improvement in recommendation ranking quality achieved through multimodal fusion, this experiment uses Normalized Discounted Cumulative Gain at 10 (NDCG@10) as the core evaluation metric. NDCG@10 measures how well the model ranks relevant items in the top-10 recommendation list. Higher NDCG@10 values indicate higher placement of highly relevant content, which better meets the requirement of prioritizing the most relevant content in educational recommendations.

For each test user, the model generates a recommendation score for the entire candidate content library and sorts the top 10 items in descending order of score. Because each sample in the dataset corresponds to a clear positive interaction item (i.e., the next clicked content), this sample is marked as highly relevant (with a relevance score of 1). The remaining non-interacting items are considered irrelevant (with a score of 0). The NDCG@10 value of the ranked list is calculated for each user and then averaged across all users to obtain the overall NDCG@10 of the model. A comparison of the NDCG@10 values of the various models is shown in Figure 4.

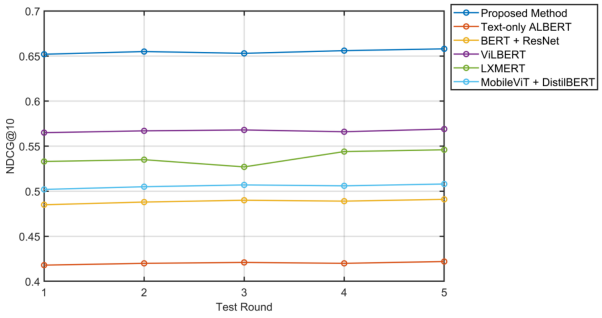

FIGURE 4. Comparison of NDCG@10 ranking performance of each model in 5 rounds of testing

Figure 4 shows that the NDCG@10 of the proposed method consistently remains above 0.65, reaching an average of 0.655. This is higher than text-only ALBERT (average 0.42) and MobileViT + DistilBERT (average 0.506), and outperforms multimodal models such as ViLBERT (average 0.567) and LXMERT (average 0.537). The results indicate that the proposed model achieves superior relevance ranking in the ranking of recommendation results, particularly in prioritizing content with high age-appropriateness and high safety value.

E. MODEL SIZE AND PEAK MEMORY USAGE

To verify the feasibility of model deployment on kindergarten edge devices, this experiment rigorously measures the disk usage of the complete inference system. The specific process is as follows. First, the trained model weights are exported, including all trainable parameters of the lightweight ALBERT encoder, ResNet-18, WaveNet audio encoder, cross-modal projection layer, gated fusion MLP, and dual-tower recommendation head. Next, all components required for inference are integrated: preprocessing scripts, TorchScript serialized models, dependency lists, and a lightweight inference engine. Peak memory usage during inference for each model is also recorded.

All files are packaged into a single compressed archive. The actual disk usage after decompression is calculated in megabytes on Ubuntu 18.04. This size represents the minimum storage space required for model deployment on resource-constrained devices such as the Jetson Nano. Table 1 shows the model size and peak memory usage.

TABLE 1. Model volume and peak memory usage

Model Name	Model Size (MB)	Peak Memory Usage (MB)	FLOPs (G)	Parameter Count (Millions)
Proposed Method	87.3	623	2.8	31.2
Text-only ALBERT	42.1	319	1.1	11.7
BERT + ResNet-50	186.5	982	12.6	139.4
ViLBERT	1228	2851	18.3	245.8
LXMERT	1536	3127	22.7	320.6
MobileViT + DistilBERT	112.4	743	4.2	39.5

Table 1 shows that the proposed lightweight multimodal model has an overall size of 87.3 MB and a peak memory usage of 623 MB, significantly smaller than ViLBERT (1228 MB, 2851 MB) and LXMERT (1536 MB, 3127 MB), and also outperforms MobileViT + DistilBERT (112.4 MB, 743 MB). The embedding layers of the are frozen during training to prevent overfitting on the relatively small dataset and to reduce parameter updates. This strategy also lowers memory consumption by limiting the dimensionality of activated tensors during forward inference. In contrast, Text-only ALBERT, which processes only text, has the smallest size (42.1 MB) and the lowest memory usage (319 MB), but sacrifices multimodal expressiveness.

F. CHILD AGE MATCHING

To assess the age-appropriateness of recommended content from a developmental psychology perspective, this experiment employs subjective quality verification using manual expert scoring. The specific implementation process is as follows: 200 samples are randomly selected from the test set recommendations, covering two age groups (100 each) for 3–4 and 5–6 years old, ensuring a balanced range of content (including topics such as traffic safety, fire and electrical safety, abduction prevention, and hygiene habits).

Three nationally certified early childhood educators with at least five years of frontline teaching experience are invited to serve as expert reviewers, independently scoring each recommendation.

Scoring is based on five criteria: cognitive adaptation, language appropriateness, visual appeal, auditory guidance, and behavioral stimulation. The scoring scale is: 1 indicates the content is completely inappropriate for the comprehension capabilities or interests of this age group (e.g., recommending an abstract rules-based text to a 3-year-old); 2 indicates that the content is partially beyond the scope or lacks appeal; 3 indicates that the content is generally appropriate but lacks significant advantages; 4 indicates that the content, format, and difficulty level are highly consistent; 5 indicates that the content is not only age-appropriate but also effectively stimulates safety awareness and behavioral emulation in children of this age group. Prior to scoring, experts conduct unified training and provide typical examples to calibrate the scoring scale and ensure consistency.

Each expert blindly evaluates all 200 samples (hiding the model source and age label). The final score for each sample is the arithmetic mean of the three experts' scores. The final age-matching index is the overall mean of the average scores of the 200 samples. Figure 5 shows the Children's age-matching score.

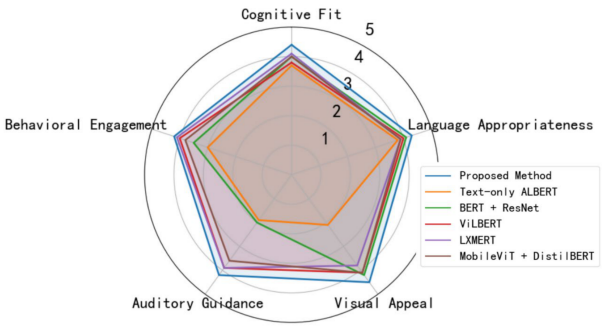

FIGURE 5. Children's age-matching score

Figure 5 shows that the proposed model achieves an average score of 4.32, outperforming Text-only ALBERT (2.9), MobileViT + DistilBERT (3.88), ViLBERT (3.96), LXMERT (3.96), and BERT + ResNet (3.56). Multimodal fusion enables the model to more accurately identify the developmental suitability of content, while Text-only ALBERT, which only uses text as a modality, scores lower.

IV. CONCLUSION

This paper proposes a lightweight multimodal recommendation method that integrates text, images, and speech for kindergarten safety education scenarios, a strategy of information fusion that also has powerful analogues in the textile field for integrating fiber data, visual patterns, and mechanical testing results for quality control. By freezing the ALBERT-base-v2 embedding layers and fine-tuning only the last two Transformer encoder layers, an efficient text encoder is constructed. A lightweight ResNet-18 and WaveNet are then combined to extract visual and audio features, respectively. Trimodal semantic alignment is achieved through 128-dimensional linear projection. A low-overhead gating mechanism is applied to dynamically fuse multimodal information. Finally, a dual-tower architecture and BPR loss are used to achieve personalized recommendations. Experiments show that this method achieves an average inference latency of 86.67 ms, a model size of 87.3 MB, an average HR@5 score of 91.4%, and an average age-appropriateness score of 4.32, outperforming all baseline models. While ensuring the feasibility of edge deployment, it effectively improves the multimodal expressiveness of safety education content and its age-appropriateness, achieving optimized performance suitable for practical applications, much like deploying machine learning models for real-time fabric defect classification in textile production.

AUTHOR CONTRIBUTIONS

Conceptualization – Ping Zhang and Xiu Fan; methodology – Ping Zhang and Xiu Fan; investigation – Ping Zhang and Xiu Fan; writing-original draft preparation – Ping Zhang and Xiu Fan. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by,

Fund Project: Scientific Research Program Funded by Education Department of Shaanxi Provincial Government

Project Name: A Practice Research on Life Education in Kindergarten Safety Management (Program No. 23JK0147)

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Xi'an Traffic Engineering Institute, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] L. S. PEK, R. W. M. MEE, W. Y. VON, M. R. ISMAIL, A. B. D. GHANI KHATIPAH, and T. S. T. SHAHDAN, "Be safe or be sorry: A scoping review on children's safety and security awareness," Quantum Journal of Social Sciences and Humanities, vol. 3, no. 5, pp. 14-25, 2022, doi: 10.55197/qjssh.v3i5.170.

[2] C. Fosu-Ayarkwah, G. Gyeabour Fosu, and I. Awortwe, "Effects of teachers' supervision on the safety of kindergarten pupils in the central region of Ghana," Open Journal of Educational Research, vol. 2, no. 6, pp. 355-366, 2022, doi: 10.31586/ojer.2022.542.

[3] N. Shaari, N. F. paiman, Y. ahmad, A. H. A. abd ghani, H. makhpol, A. M. radzi, et al., "Child Occupant Safety Program: A Study at Selected Kindergartens in Kajang and Putrajaya," Journal of the Society of Automotive Engineers Malaysia, vol. 9, no. 1, pp. 23-28, 2025, doi: 10.56381/jsaem.v2i1.75.

[4] Y. Yuan, Z. Li, and B. Zhao, "A survey of multimodal learning: Methods, applications, and future," ACM Computing Surveys, vol. 57, no. 7, pp. 1-34, 2025, doi: 10.1145/3713070.
[5] J. Moon, S. Yeo, Banihashem Sk, and O. Noroozi, "Using multimodal learning analytics as a formative assessment tool: Exploring collaborative dynamics in mathematics teacher education," Journal of Computer Assisted Learning, vol. 40, no. 6, pp. 2753-2771, 2024, doi: 10.1111/jcal.13028.
[6] A. Yusuf, N. M. Noor, and S. Bello, "Using multimodal learning analytics to model students' learning behavior in animated programming classroom," Education and Information Technologies, vol. 29, no. 6, pp. 6947-6990, 2024, doi: 10.1007/s10639-023-12079-8.
[7] X. Wang, G. Chen, G. Qian, P. Gao, X. Y. Wei, and Y. Wang, "Large-scale multi-modal pre-trained models: A comprehensive survey," Machine Intelligence Research, vol. 20, no. 4, pp. 447-482, 2023, doi: 10.1007/s11633-022-1410-8.
[8] X. Han, Y. T. Wang, J. L. Feng, C. Deng, Z. H. Chen, Y. A. Huang, et al., "A survey of transformer-based multimodal pre-trained modals," Neuro-computing, vol. 515, no. 1, pp. 89-106, 2023, doi: 10.1016/j.neucom.2022.09.136.
[9] H. I. Liu, M. Galindo, H. Xie, L. K. Wong, H. H. Shuai, Y. H. Li, et al., "Lightweight deep learning for resource-constrained environments: A sur-vey," ACM Computing Surveys, vol. 56, no. 10, pp. 1-42, 2024, doi: 10.1145/3657282.
[10] M. Ponti, "Screen time and preschool children: Promoting health and development in a digital world," Paediatrics & Child Health, vol. 28, no. 3, pp. 184-192, 2023, doi: 10.1093/pch/pxac125.
[11] M. Oybarchin, "Psychological and pedagogical aspects of the development of visual activities of preschool children," qo 'qon universiteti xabarnomasi, vol. 8, no. 1, pp. 101-105, 2023, doi: 10.54613/ku.v8i8.815.
[12] O. V. Zashchirinskaia, "Nonverbal patterns of preschooler's perception of visual images with the help of eye-tracker method usage," Current psychology, vol. 40, no. 1, pp. 442-453, 2021, doi: 10.1007/s12144-018-9960-1.
[13] G. De Mello, M. N. A. Ibrahim, N. Arumugam, M. S. Husin, N. H. O. Ma'mor, and S. Dharinee, "Nursery rhymes: Its effectiveness in teaching of English among pre-schoolers," International Journal of Academic Research in Business and Social Sciences, vol. 12, no. 6, pp. 1914-1924, 2022, doi: 10.6007/IJARBSS/v12-i6/14124.
[14] T. N. Fitria, "Using nursery rhymes in teaching English for young learners at childhood education," Athena: Journal of Social, Culture and Society, vol. 1, no. 2, pp. 58-66, 2023, doi: 10.58905/athena.v1i2.28.
[15] Y. Christina and P. Pujiarto, "The Effectiveness of Nursery Rhymes Media to Improve English Vocabulary and Confidence of Children (4-5 Years) in Tutor Time Kindergarten," Journal of Education Research, vol. 4, no. 3, pp. 1326-1333, 2023, doi: 10.37985/jer.v4i3.406.
[16] F. Hayat, "The effect of education using video animation on elementary school in hand washing skill," Acitya: Journal of Teaching and Education, vol. 3, no. 1, pp. 44-53, 2021, doi: 10.30650/ajte.v3i1.2135.
[17] M. Tuo and B. Long, "Construction and application of a human-computer collaborative multimodal practice teaching model for preschool education," Computational Intelligence and Neuroscience, vol. 2022, no. 1, Art. no. 2973954, 2022, doi: 10.1155/2022/2973954.
[18] E. Fundelius, T. Wade, A. Robbins, S. Wang, M. A. McConomy, and K. Fumero, "Universal design principles for multimodal representation in litera-cy activities for preschoolers," Inclusive Practices, vol. 2, no. 1, pp. 13-21, 2023, doi: 10.1177/27324745221140380.
[19] F. Ally, J. D. Pillay, and N. Govender, "Teaching and learning considerations during the COVID-19 pandemic: Supporting multimodal student learning preferences," African Journal of Health Professions Education, vol. 14, no. 1, pp. 13-16, 2022, doi: 10.7196/AJHPE.2022.v14i1.1468.
[20] Q. Si, T. S. Hodges, and J. M. Coleman, "Multimodal literacies classroom instruction for K-12 students: a review of research," Literacy Research and Instruction, vol. 61, no. 3, pp. 276-297, 2022, doi: 10.1080/19388071.2021.2008555.
[21] X. Ren, W. Yang, X. Jiang, G. Jin, and Y. Yu, "A deep learning framework for multimodal course recommendation based on LSTM+ attention," Sustainability, vol. 14, no. 5, pp. 2907, 2022, doi: 10.3390/su14052907.
[22] M. Ryu, G. Lee, and K. Lee, "Knowledge distillation for bert unsupervised domain adaptation," Knowledge and Information Systems, vol. 64, no. 11, pp. 3113-3128, 2022, doi: 10.1007/s10115-022-01736-y.
[23] I. Cho and U. Kang, "Pea-KD: Parameter-efficient and accurate Knowledge Distillation on BERT," Plos One, vol. 17, no. 2, Art. no. e0263592, 2022, doi: 10.1371/journal.pone.0263592.
[24] S. Safitri, M. Alii, and O. Mahmud, "Murottal Audio as a Medium for Memorizing the Qur'an in Super-Active Chil-dren," Journal International Inspire Education Technology, vol. 1, no. 2, pp. 111-124, 2022, doi: 10.55849/jiiet.v1i2.87.
[25] S. K. Uppada, P. Patel, and S. B, "An image and text-based multimodal model for detecting fake news in OSN's," Journal of Intelligent Information Systems, vol. 61, no. 2, pp. 367-393, 2023, doi: 10.1007/s10844-022-00764-y.
[26] A. Chen, S. Ambrogio, P. Narayanan, A. Okazaki, C. Mackin, A. Fasoli, et al., "Demonstration of transformerbased ALBERT model on a 14nm analog AI inference chip," Nature Communications, vol. 16, no. 1, pp. 8661, 2025, doi: 10.1038/s41467-025-63794-4.
[27] M. S. Sidhu, N. A. A. Latib, and K. K. Sidhu, "MFCC in audio signal processing for voice disorder: a review," Multimedia Tools and Applications, vol. 84, no. 10, pp. 8015-8035, 2025, doi: 10.1007/s11042-024-19253-1.
[28] A. I. Siam, A. A. Elazm, N. A. El-Bahnasawy, G. M. El Banby, and F. E. Abd El-Samie, "PPG-based human identification using Mel-frequency cepstral coefficients and neural networks," Multimedia Tools and Applications, vol. 80, no. 17, pp. 26001-26019, 2021, doi: 10.1007/s11042-021-10781-8.
[29] S. Mehra, V. Ranga, and R. Agarwal, "Multimodal Integration of Mel Spectrograms and Text Transcripts for Enhanced Automatic Speech Recognition: Leveraging Extractive Transformer-Based Approaches and Late Fusion Strategies," Computational Intelligence, vol. 40, no. 6, Art. no. e70012, 2024, doi: 10.1111/coin.70012.
[30] U. R. Pol, P. S. Vadar, and T. T. Moharekar, "Hugging Face: Revolutionizing AI and NLP," International Journal for Research in Applied Science and Engineering Technology, vol. 12, no. 8, pp. 1121-1124, 2024, doi: 10.22214/ijraset.2024.64023.
[31] A. Ajibode, A. A. Bangash, F. R. Cogo, B. Adams, and A. E. Hassan, "Towards semantic versioning of open pre-trained language mod-el releases on hugging face," Empirical Software Engineering, vol. 30, no. 3, pp. 1-63, 2025, doi: 10.1007/s10664-025-10631-3.
[32] Z. Yin, E. Xing, and Z. Shen, "Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective," Advances in Neural Information Processing Systems, vol. 36, no. 1, pp. 73582-73603, 2023, doi: 10.52202/075280-3219.
[33] A. Ullah, H. Elahi, Z. Sun, A. Khatoon, and I. Ahmad, "Comparative analysis of AlexNet, ResNet18 and SqueezeNet with diverse modification and arduous implementation," Arabian Journal for Science and Engineering, vol. 47, no. 2, pp. 2397-2417, 2022, doi: 10.1007/s13369-021-06182-6.
[34] R. Helaly, S. Messaoud, S. Bouaafia, M. A. Hajjaji, and A. Mtibaa, "DTL-I-ResNet18: facial emotion recognition based on deep trans-fer learning and improved ResNet18," Signal, Image and Video Processing, vol. 17, no. 6, pp. 2731-2744, 2023, doi: 10.1007/s11760-023-02490-6.
[35] A. Paul, Z. Wu, K. Liu, and S. Gong, "Robust multi-objective visual bayesian personalized ranking for multimedia recommendation," Applied Intelligence, vol. 52, no. 4, pp. 3499-3510, 2022, doi: 10.1007/s10489-021-02355-w.