Application of Multimodal Generation Model in Short Video Content Personalized Generation

Minghui Yang
New Media E-commerce Institute, Chongqing Institute of engineering, Chongqing 400000, Chongqing, China
E-mail : minghuiyangminghui@163.com

Keywords: personalized content, multimodal generation, stochastic paint optimizer with intelligent convolutional neural network (SPO-IntelliConvNet), modalities

Received: June 23, 2025

The rise of short video platforms has led to a higher demand for rapidly generated personalized content. Existing systems either struggle with high levels of customization or require large amounts of data, limiting real-time production. A multimodal generation model serves as the focus of study to generate customized short video content that adapts to user preferences as well as their behavioral patterns. The objective targets an integrative model using text alongside image and audio data to make context-specific short video content, which delivers personalized entertainment. First, it analyses user preferences from interaction data and then synthesizes corresponding video content using a novel method called a stochastic paint optimizer with an intelligent convolutional neural network (SPO-IntelliConvNet). The SPO component ensures optimal representation of multimodal content by improving feature selection and parameter tuning through stochastic search algorithms modelled after the dynamics of abstract paintings. The IntelliConvNet is used to combine and interpret several modalities, allowing for efficient personalization that is consistent with user preferences. To develop personalized content, user preference data is collected, which includes interactions such as video views and comments. The model employs natural language processing (NLP), audio processing, and computer vision to merge text, image, and audio modalities. Pre-processing includes tokenization for text, Canny edge detection for images, and Wiener filtering for audio, optimizing each modality for better analysis and feature extraction using principal component analysis (PCA) to reduce the dimensions of features from all three modalities to lower dimensions while preserving essential information. This proposed approach achieved superior personalized content development, leading to increased user satisfaction and engagement. The performance of the proposed method was evaluated using BLEU-4 (0.55), ROUGE-L (0.79), METEOR (0.72), and CIDEr (0.80). The system's ability to successfully incorporate multimodal data resulted in more precise video customization, as demonstrated by interaction metrics and user comments. This multimodal generation model provides an advanced solution for creating personalized short video content, increasing the user experience with highly tailored content.

Povzetek: Študija predlaga večmodalni generativni model za osebno prilagojene kratke videe, ki se z združitvijo NLP, računalniškega vida in zvočne obdelave ter novim optimizatorjem SPO-IntelliConvNet iz uporabniških interakcij uči preference in sintezira kontekstno ustrezno vsebino v realnem času.

1 Introduction

The sudden growth of short video platforms has resulted in an increasing demand for customized content that boosts user interaction. With the widespread use of mobile technology and social media, short videos have emerged as a prominent medium of communication and entertainment. Users anticipate content that is personalized to their interests, and therefore, personalization becomes a central theme of video creation. Traditional metadata-based user tracking recommendation systems are likely to overlook the multimodal nature of short videos and therefore make less interactive and less relevant recommendations [1]. Dynamic, engaging content results from short video personalization, which processes different data types through text, images, and audio. Textual analysis of captions and visual analysis of object detection and scene understanding, and audio examination of speech and ambient noise create content that matches user attention patterns [2].

Real-time video generation through customized content becomes possible because the system avoids incorporating pre-selected templates to suit individual needs. This provides for real-time content generation,

which responds to users' specific behaviour, leading to better storytelling and interaction. The behaviours of user interaction (likes, shares, and comments) are integrated with the content of videos and customized according to the feedback presented instantly, and this makes the content of the video more meaningful and engaging [3]. Videos filmed by individuals now proliferate their influence in entertainment, educational, and marketing sectors, as well as improve social media engagement. The personalization of content enables users to achieve maximum educational potential in videos since the content generation method creates true brand interactions, and video content generates precise information that suits the demands of the viewer [4]. The generation of personalized short videos intends to generate video experiences that are engaging and adaptive depending on user interactions and interests. This technique uses text, images, and audio to dynamically adjust content and make it more relevant and engaging to the viewer. It is regularly applied to entertainment, education, and advertising, providing users with customized and engaging video experiences to improve storytelling and viewer engagement [5].

The objective is to construct a multimodal generation model for individualized short video content with integration of text alongside images, together with audio features. The model called SPO-IntelliConvNet improves feature extraction to enhance content customization and increase user engagement.

1.1 Significance of the contribution

• The system enhances personalization through both user preferences and adaptable content creation methods in short video generation.

• Adaptive multimodal feature fusion with NLP, vision, and audio – The framework jointly processes text (NLP), images (computer vision), and audio (signal processing) using modality-specific preprocessing: tokenization and embedding for text, Canny edge detection for images, and Wiener filtering for audio. NLP plays a central role by extracting semantic meaning, sentiment, and contextual cues from user comments and video descriptions, which guide the personalization process.

• Feature extraction using Principal Component Analysis (PCA) enhances content synthesis by identifying key personalization patterns.

• Novel integration of optimization and deep learning proposes SPO-IntelliConvNet, the first framework to integrate the Stochastic Paint Optimizer (SPO) with an adaptive convolutional neural network for personalized short video generation. Unlike conventional multimodal models that rely on static parameters, SPO dynamically adjusts convolutional filters, fusion weights, and hyperparameters, enabling adaptive multimodal learning.

• Comprehensive evaluation of personalization quality – Beyond automated metrics (BLEU, ROUGE-L, METEOR, CIDEr), we incorporate baseline comparisons, statistical significance testing, and user-centric evaluation (e.g., watch-time analysis and subjective preference ratings). This multi-layered validation ensures that performance gains are not only statistically meaningful but also translate into measurable improvements in user engagement and satisfaction.

1.2 Research organization

The research follows the structure outlined below. Section 2 presents a review of the literature on multimodal content generation. Section 3 explains the proposed SPO-IntelliConvNet model and its methodology for personalized short video generation. Section 4 details the experimental results, highlighting the model's performance based on evaluation metrics. Finally, Section 5 describes the conclusion.

2 Literature review

A user behavior-aware multi-task learning model for enhanced short video recommendation (UBA-SVR) was presented in [6] by utilizing knowledge about changing user communication. The outcomes indicated that the suggested approach significantly improves several prediction tasks instantly.

A multimodal short video recommendation framework was proposed, integrating text, image, and audio processing [7]. Crawler technology extracted text, frame processing captured images, and voice data were analyzed. Another data fusion approach transformed video features into a dense space to enhance the recommendation process. While the model provided a better understanding of user behavior, it encountered difficulties with improving both multimedia interactions and computation speed optimization.

A unique recommendation strategy for short video platforms was proposed in [8]. The suggested approach combines an error back propagation neural network (EBPNN) with the term frequency inverse document frequency (TF-IDF) algorithm to investigate the possible relationship between users and videos for text mining. The findings revealed that the model's forecasting accuracy had increased substantially, with scores of 73.50% and 88%.

A knowledge graph-based recommendation system for short new media videos was proposed in [9]. The findings demonstrated that using knowledge graphs to recommend short videos could significantly increase the standard of content suggestions and provide consumers with a more engaging and customized viewing experience. Table 1 shows the description of related works.

Table 1: Comparison between the SOTA method in short video content personalized generation

Reference	**Dataset**	**Methodology**	**Key findings**
Zhu (2025) [10]	Short video platform dataset	Deep Learning and Reinforcement Learning (DLRL)	Achieved click-through rate (6.22%), Accuracy (0.877), and Recall (0.858)
Yang (2025) [11]	Short video dataset	Deep Learning-based content analysis and recommendation	Improved average precision by 4.7%, 3.3%, and 4.0%, and recall by 3.5%, 1.2%, and 2.1% compared to CF, MF, and Content-Based Filtering
Lu and Nam (2021) [12]	Short video dataset	Artificial Intelligence–based Video Auditing	Optimized content screening, enhanced original video support, and ensured sustainable, healthy development of mobile short video platforms.
Song and Liu (2024) [13]	TikTok dataset	Key and Objective Indicator System Based Communication Network (KOIS-PN) communication network + Attention-based Bidirectional Long Short-Term Memory Network (AT-BiLSTM) neural network	The KOISPN-ATBiLSTM short video model for international communication networks demonstrates certain growth and validity.

2.1 Problem statements

Short video platforms require accurate personalized recommendation systems to enhance user engagement. Existing methods, such as UBA-SVR [6], multimodal frameworks [7], EBPNN-TF-IDF hybrids [8], and knowledge graph-based systems [9], improve prediction accuracy but face limitations in handling multimodal interactions, semantic relationships, and computational efficiency. The *SPO-IntelliConvNet* integrates user behavior modeling, multimodal feature extraction, and knowledge graphs to overcome these gaps, providing faster, context-aware, and highly personalized recommendations. This approach combines prior strengths while mitigating their respective limitations.

3 Methodology

The proposed methodology integrates SPO-IntelliConvNet to enhance personalized short video generation. Including user data is pre-processed, and feature extraction is performed using PCA. IntelliConvNet identifies hierarchical patterns in multimodal content, while SPO optimizes model parameters for improved accuracy, enabling adaptive and AI-driven video synthesis. Figure 1 illustrates the framework of the SPO-IntelliConvNet model.

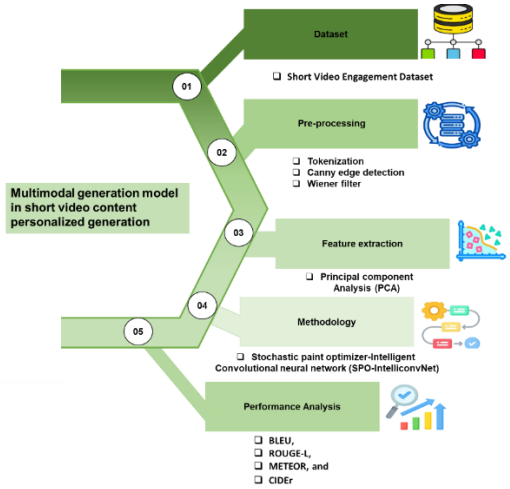

Figure 1: The framework of the proposed model

3.1 Data set

The Short Video Engagement Dataset (https://www.kaggle.com/datasets/programmer3/short-video-engagement-dataset) was gathered from Kaggle; it

includes YouTube Shorts, TikTok, and Instagram Reels, capturing user engagement and multimodal features. It represents video content data, including user behaviour (views, likes, comments, and shares), text features (title and description length), image attributes (edge intensity and colour histogram), and audio properties (spectral entropy, audio intensity). The engagement score (0 for low, 1 for high) serves as the target variable, helping analyze patterns in personalized short video content consumption. To clearly state its size (17,654 rows, 12 columns), the modalities covered (user interaction, textual, visual, and audio features), and the target variable (binary engagement score). We also now explicitly acknowledge its limitations in diversity and representativeness, noting that while it includes data from three major platforms (YouTube Shorts, TikTok, Instagram Reels), it may not fully capture domain-specific or regional variations in short video content.

3.2 Pre-processing

Pre-processing enhances multimodal data for personalized short video generation, applying tokenization for text, Canny edge detection for images, and Wiener filtering for audio to optimize feature extraction and content synthesis. Audio processing, computer vision, and NLP characteristics are combined in an organized multimodal fusion approach. The fusion improves semantic coherence and contextual significance by capturing relationships between modalities. The model creates information by coordinating textual, visual, and audio data, which shows a thorough comprehension of the user's intent and preferences. This approach guarantees that every input modality makes a valuable contribution to personalization, enhancing user engagement metrics and overall content quality.

3.2.1 Tokenization

The Short Video Engagement Dataset is used for the tokenization process, resulting in a structured and clean text representation, improving its efficiency. By removing noise, stop words, and inconsistencies, the processed text enhances semantic understanding, ensuring more relevant and personalized content generation. This structured tokenized data leads to improved classification accuracy, better content personalization, and enhanced user engagement in short video recommendations.

3.2.2 Canny edge detection

This technique requires the Short Video Engagement Dataset to extract useful edge features from images, which enhances its operational efficiency. The method identifies sudden intensity changes, assisting in the detection of object boundaries and eliminating redundant data. By using Gaussian filtering to blur the image, followed by gradient calculation and non-maximum suppression, it efficiently retains important edges. Double thresholding and edge tracking then refine the outcome, retaining only strong edges. This technique supports the feature extraction process by limiting computational complexity without compromising the accuracy of generating customized short video content. The Gaussian filtering is represented in equation (1).

$$M(p,q) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{p^2+q^2}{2\sigma^2}\right) \quad (1)$$

Where $M(p,q)$ indicates the Gaussian filter value at pixel position, $\exp\left(-\frac{p^2+q^2}{2\sigma^2}\right)$ signifies the exponential decay function, p, q represent the filter coordinates relative to the center of the fixed kernel, π indicates the balance of pixel intensities, and σ signifies the standard deviation of the Gaussian distribution. After noise reduction, image gradients are computed using Sobel operators to detect edge intensity variations. The gradient magnitude and direction are given by equations (2 and 3).

$$M = \sqrt{M_p^2 + M_q^2} \quad (2)$$

$$\theta = \tan^{-1}\frac{M_q}{M_p} \quad (3)$$

Non-maximum suppression thins edges, double thresholding classifies strong and weak edges, and hysteresis tracking retains weak edges linked to strong ones. Where M indicates the magnitude of the personalized video feature vector, θ represents the gradient direction, M_p signifies the primary feature magnitude, and M_q indicates the secondary feature magnitude. Canny edge detection serves as a crucial pre-processing step to enhance image analysis accuracy. Refining edge structures ensures robust performance in subsequent image-processing tasks.

3.2.3 Wiener filtering

The Short Video Engagement Dataset data are used for this method, as it performs the greatest trade-off between inverse filtering and noise attenuation from audio signals, enhancing its efficiency. It efficiently subtracts additive noise without affecting useful signal components and improves the total quality. It is the best optimum in the reduction of mean square error (MSE) that ensures better reconstruction of the signal. In the pre-processing phase, Wiener filtering is employed for suppressing background noise and recovering original audio features to enhance clarity and accuracy in audio-based applications, as derived in equation (4).

$$W(v) = \frac{P_m(v)}{P_m(v)+P_n(v)} \quad (4)$$

Where $W(v)$ signifies the personalization weight of video v, $P_m(v)$ indicates the probability or score that video v matches the user interests, and $P_n(v)$ represents the probability or score that video v does not match the user interests.

Pre-processing enhances data quality for feature extraction. Tokenization refines textual data by segmenting meaningful units. Canny edge detection enhances the clarity of image structures, and it sharpens object boundaries. Wiener filtering is an approach to audio denoising to preserve the quality of the signal. The enhancement of these outputs through design produces more accurate features, which enhance the accuracy of the models used in the extraction procedure.

3.3 Feature extraction

Principal component analysis (PCA)

PCA is a dimensionality reduction technique of representing high dimensional data into a lower dimensional space and retaining critical information. It identifies main components (PCs) which can be measured with maximum variance, which minimises redundancy and noise. PCA improves the efficiency of the computations and it is extensively applied in the areas of feature extraction, pattern recognition, and machine learning (ML).

PCA is used as a feature extraction method to minimize dimensions without losing critical data. With a collection of transformed, pre-processed data, PCA is used to determine the most important components that explain maximum variance. $W = W_1, W_2, W_o$ represent the feature set obtained from preprocessing, where the covariance matrix C and eigenvalues $1 > \lambda 2 > \cdots . > \lambda p$. The PCs are calculated as in equations (5 and 6).

$$Z_1 = b_1' W = b_{11} W_1 + b_{12} W_2 + \cdots . b_{1o} W_o \quad (5)$$

where:

Z_1 is the first principal component,

b_1' is the eigenvector corresponding to the largest eigenvalue,

b_{11}: contribution of feature W_1 to Z_1,

b_{12}: contribution of feature W_2 to Z_1,

Let $W = W_1, W_2, W_o$ be the random variables, with covariance matrix C and eigenvalues $\lambda 1 > \lambda 2 > \cdots . > \lambda p$.

$$Z_2 = b_2' W = b_{21} W_1 + b_{22} W_2 + \cdots . b_{2o} W_o \quad (6)$$

$$\vdots \quad \vdots$$

$$Z_O = b_O' W = b_{O1} W_1 + b_{O2} W_2 + \cdots . b_{OO} W_O \quad (7)$$

Variance $Var(Z_j)$, covariance $Cov(Z_j, Z_l)$ in equations (7 to 9).

$$Var(Z_j) = b_j' \sum b_j; j = 1,2, \ldots, o \quad (8)$$

$$Cov(Z_j, Z_l) = b_j' \sum b_l; j = 1,2, \ldots, o \quad (9)$$

The process of extracting features, improves the representation of data by making it less redundant whilst it preserves important information gained during pre-processing stages. j Index of the principal component (ranging from 1 to o, where o is the number of principal components retained). This boosts computation efficiency, removes redundancy and boosts accuracy of classification allowing better data representation to optimize model performance in short video content personalized generation.

3.4 Personalized content generation using stochastic paint optimizer-tuned intelligent convolutional neural network (SPO-IntelliConvNet)

The integration of SPO within IntelliConvNet is designed as an optimization-driven enhancement. SPO functions as a parameter-tuning mechanism, where convolutional filters, feature fusion weights, and learning rates are iteratively optimized using stochastic search principles. IntelliConvNet then operates on these optimized parameters to extract, align, and combine modality-specific features. This synergy ensures that each modality, text, audio, and image contribute meaningfully to the final representation. The result is a CNN variant that self-adjusts its internal representations, making the integration explicit and functionally distinct from standard architectures. The Stochastic Paint Optimizer with IntelliConvNet to achieve adaptive multimodal feature extraction. Unlike traditional models, SPO dynamically optimizes convolutional filters and fusion weights, enabling context-aware learning across text, image, and audio modalities. The creative analogy with color theory highlights complementary interactions, inspiring new perspectives in multimodal optimization-driven content generation for personalized short video applications.

3.4.1 Intelligent convolutional neural network (IntelliConvNet)

The deep learning (DL) structure known as CNN exists to process structured grid data, specifically images. The distinct layers of CNNs include convolutional and pooling, and fully connected sections to extract spatial data while building hierarchical patterns from this data. By utilizing weight-sharing mechanisms, CNNs reduce computational complexity while improving feature learning. These models provide multiple applications in visual object identification alongside pattern recognition and identification of objects in images.

The implementation of IntelliConvNet technology allows the multimodal generation model to create customized short video content. Advanced CNN mechanisms in the model transform text and image together with audio features for enhancement purposes. IntelliConvNet applies adaptive convolutional filtering with attention mechanisms and optimization techniques to obtain improved efficiency and accuracy. The model dynamically chooses dynamic kernels, which allows it to adjust convolutional parameters according to content features for high-quality feature representation.

PCA features two necessary operations that enable dimension reduction while preserving important information, together with feature extraction. SPO optimization allows convolutional layers to master filter weights independently from changing content complexities in a self-learning process. Depth-wise separable convolutions and residual connections are used to reduce the computational load at the expense of minimal loss in model performance. The model is also optimized using Neural Architecture Search (NAS) to learn the IntelliConvNet architectures automatically, achieving a trade-off between accuracy and computation cost. It provides personalized video content creation with maximum flexibility and multimodal learning. Also, self-supervised learning mechanisms are proposed for feature representation improvement with reduced dependency on large annotated datasets. Spatial and channel attention-based hybrid attention further promotes content-based feature learning. All of these enhance video personalization by providing precise, high-quality, and context-aware content generation. Adaptive convolution filtering and feature transformation mathematical equations would be described as needed to present the mathematical foundation of these advancements. The system identifies appropriate features to enhance individualized video synthesis using the combined work of IntelliConvNet and SPO. This combined solution optimizes video customization according to the preferences of users, which results in increased engagement and the level of user satisfaction.

3.4.2 Stochastic paint optimizer (SPO)

SPO, a metaheuristic optimization algorithm, is designed based on principles derived from color theory. It stimulates the process of painting by treating the search space as a canvas and solutions as paints. The algorithm evaluates solutions based on a beauty index, which represents the objective function value. It iteratively refines solutions using different colour combination techniques, ensuring an optimized outcome. Figure 2 shows the SPO flowchart.

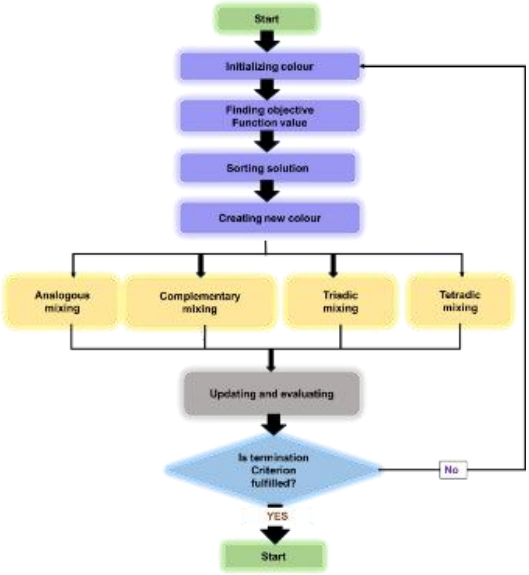

Figure 2: The flowchart of SPO

The SPO executes its operations by creating an initial population, which contains features from text, image, and audio elements serving as solutions. Then, it clusters and aggregates these attributes using various combination methods to enhance representation. The solutions are assessed based on a beauty index, which serves as the fitness function in the search. Through iterative stochastic processes, SPO optimizes solutions by dynamically fusing and evolving characteristics until achieving optimal convergence, ensuring greater precision and efficiency in multimodal content generation.

i. Analogous combination for feature selection

This technique blends three adjacent feature sets from a modality to refine content representation as explained in equation (10).

$$D_{new}^{1} = D_j + rand.(D_{j+1} - D_{j-1}) \quad (10)$$

Where D_j is the selected feature vectors, $rand$ is a randomly generated vector in [0,1], D_{j-1} indicates the feature vector of the previous adjacent segment in the video sequence, D_{j+1} indicates the feature vector of the next adjacent segment in the video sequence, and D_{new}^{1} represents the refined or augmented feature vector for the j^{th} segment.

ii. Complementary combination for feature fusion

Complementary features from different modalities (text, image, audio) are fused using equation (11).

$$D_{new}^{2} = D_j + rand\,.\,(D_O - D_S) \quad (11)$$

Where D_O is the original or target feature vector, D_S signifies the source or current feature vector, and D_{new}^{2} indicates the newly generated feature representation.

iii. Triadic combination for multimodal integration

This method integrates three distinct feature sets to maintain a balanced representation across modalities, with the function of Triadic shown in equation (12).

$$D_{new}^{3} = D_j + rand\,.\left(\frac{D_O + D_T + D_S}{3}\right) \quad (12)$$

Where D_T indicates the textual features, D_S represents the style or audio features, and D_{new}^{3} signifies the triadic integrated feature representation.

iv. Tetradic combination for final video synthesis

The final step involves blending four feature sets across modalities for optimal content synthesis using equations (13 and 14).

$$D_{new}^{4} = D_j + \left(\frac{rand_1.D_O + rand_2.D_T + rand_3.D_S + rand_4.D_{rand}}{4}\right) \quad (13)$$

$$D_{rand} = LB + rand\,.\,(UB - LB) \quad (14)$$

Where $UB\ and\ LB$ indicate the upper and lower feature bounds, $rand_1, rand_2, rand_3, rand_4$ are random vectors in [0,1] used to scale the feature vector between $UB\ and\ LB$, and D_{new}^{4} signifies the tetradic integrated feature representation.

SPO is employed to maximize multimodal short video content creation by addressing text, image, and audio features as components of paint. The algorithm incorporates these components based on various colour-mixing methods, promoting feature extraction and fusion for more customized content creation. The optimization process guarantees that synthesized videos are optimized according to user preferences, resulting in enhanced engagement and satisfaction. Figure 3 illustrates an IntelliConvNet architecture extracting hierarchical features from multimodal data (text, image, and audio), transforming them into feature maps, and classifying models into success/failure categories, enhancing personalized short video generation.

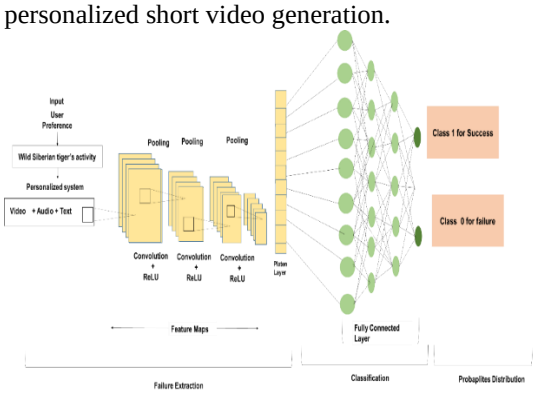

Figure 3: IntelliConvNet feature extraction for video generation

SPO-IntelliConvNet represents an integrated solution uniting SPO with IntelliConvNet to optimize personalized short video creation. SPO optimally prefers and learns multimodal features (image, text, audio) through stochastic processing, whereas IntelliConvNet augments feature learning through adaptive convolutional filtering and intelligent pooling. Algorithm 1 represents the way SPO-IntelliConvNet integrates adaptive convolutional filtering and stochastic optimization, enhancing multimodal feature extraction, video customization, and personalization while reducing computational complexity for efficient short video generation.

Algorithm 1: Stochastic Paint Optimizer-Tuned Intelligent Convolutional Neural Network (SPO-IntelliConvNet)

1. Initialize CNN weights W randomly $(-0.1\ to\ 0.1)$
2. Initialize paint parameters $P = 0.5$
3. For epoch $= 1\ to\ 50$ do
4. For each batch of 50 samples in (X, Y) do
5. Apply stochastic paint variation: $P\_new = P + random(-0.05, 0.05)$
6. Forward pass: $y\_pred = CNN(x\_batch, W)$
7. Compute loss: $L = MSE(y\_batch, y\_pred)$
8. Backpropagate and update W with learning rate 0.001
9. End For
10. Evaluate validation accuracy (batch size 50)
11. Update P using stochastic update rule: $P = P\_new$
12. End For
13. Return optimized weights $W*$ and paint parameters $P*$

4 Result and discussion

This section discusses the outcome of individualized short video content creation with the SPO-IntelliConvNet model and the traditional IntelliConvNet method; testing significant metrics is essential for determining the quality and accuracy of created content to reference content. To substantiate claims of user satisfaction, this research extends beyond automated metrics such as BLEU, ROUGE, METEOR, and CIDEr. The proposed model will be validated against IntelliConvNet baseline methods under identical experimental settings, with performance improvement methods trained in the Short Video Engagement Dataset. The research was conducted on a Windows 10 computer with an Intel i5 processor and 16 GB of RAM for data processing. Python 3.12.1 was used as the programming language, permitting the execution of the proposed SPO-IntelliConvNet model for multimodal generation in personalized short video content creation (Computing Method). Table 2 presents key metric values of the proposed model SPO-IntelliConvNet.

Table 2: Performance metrics of SPO-IntelliConvNet

Metrics	**IntelliConvNet**	**SPO-IntelliConvNet [Proposed]**
BLEU-1	0.69	0.75
BLEU-2	0.59	0.65
BLEU-3	0.54	0.60
BLEU-4	0.49	0.55
METEOR	0.66	0.72
CIDEr	0.72	0.80
ROUGE-1	0.76	0.82
ROUGE-2	0.69	0.75
ROUGE-L	0.73	0.79

4.1 Bilingual evaluation understudy (BLEU)

Measures textual overlap, ensuring generated video captions align syntactically with user preferences for accurate personalization. It evaluates the similarity between reference and generated text using n-gram precision. It evaluates the syntactic and lexical similarity. This mathematical function is as follows in equations (15 and 16).

$$\text{BLEU} = \text{BP}.\exp\left(\sum_{n=1}^{m} g_n \log s_n\right) \tag{15}$$

$$BP = \min\left(1,\ e^{1-\frac{k}{b}}\right) \tag{16}$$

The brevity penalty (BP) accounts for the difference between the generated text length and the reference length to penalize overly short outputs. Reference length(k), generated text length(b), m refers to the **maximum n-gram order** considered, n −gram precision (s_n), weight of n-gram precision (g_n). In short video content generation, BLEU helps evaluate how closely generated video captions (from IntelliConvNet) match actual user-generated captions. A higher BLEU score means better textual accuracy, ensuring that the content is relevant to user preferences. Figure 4 illustrates the BLEU score outcome, highlighting model performance across different epochs.

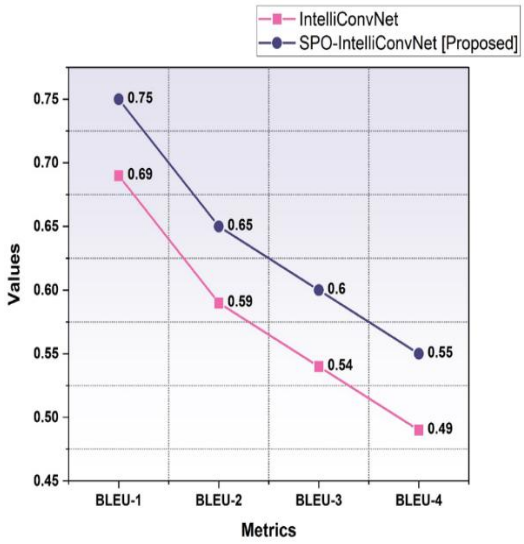

Figure 4: Visual representation of BLEU

The BLEU scores indicate the model's effectiveness in generating text aligned with reference data. The proposed SPO-IntelliConvNet has a BLEU-1 score of 0.75, shows strong unigram precision, while BLEU-2 (0.65), BLEU-3 (0.60), and BLEU-4 (0.55) reflect progressively lower scores as n-grams increase. The graph shows steady improvement across epochs, confirming enhanced linguistic accuracy. However, the traditional IntelliConvNet has the low BLEU scores of BLEU-1 (0.69), BLEU-2 (0.59), BLEU-3 (0.54), and BLEU-4 (0.49). The research outcome demonstrates the model's ability to optimize multimodal content generation effectively. The consistent BLEU score progression highlights its adaptability in refining text coherence over time.

4.2 Metric for evaluation of translation with explicit ordering (METEOR)

Evaluates semantic alignment and synonym matching, validating that generated multimodal content reflects the contextual meaning of user intent. It improves BLEU by incorporating synonyms, stemming, and word arrangement. It assesses both precision and recall instead of relying solely on n-gram matching to enhance effectiveness in capturing semantic meaning expressed in equations (17 and 18).

$$\text{METEOR} = \text{D}_{mean} \times (1 - penalty) \quad (17)$$

$$\text{D}_{mean} = \frac{S \times Q}{\alpha S + (1-\alpha) Q} \quad (18)$$

Where D_{mean} represents the harmonic mean of precision (P) and recall (R) with a tunable parameter α. Where a fraction of matched words in the generated text is represented as (S), a fraction of matched words in the reference text as (Q), and the Weighting factor as (α).

Since personalized video generation relies on NLP for user preferences, METEOR ensures that generated captions are not just lexically but also semantically accurate. It improves personalization by capturing meaning variations, like synonyms and different word orders, which are common in user interactions. Figure 5 illustrates the METEOR score.

The conventional IntelliConvNet has a low METEOR score of 0.66, whereas the suggested SPO-IntelliConvNet has a METEOR score of 0.72, demonstrating the model's strong ability to generate personalized short video content by effectively incorporating multimodal data (text, image, and audio). The SPO-IntelliConvNet method optimizes user preferences, ensuring enhanced synonym matching and stemming. The steady increase in METEOR across epochs confirms the model's progressive learning, improving content relevance and coherence for a highly engaging user experience.

4.3 Consensus-based image description evaluation (CIDEr)

Assesses consensus with multiple human references, validating descriptive quality and alignment of generated video content with user-specific expectations. It evaluates image captioning by assessing the alignment of generated text with multiple reference captions. It places higher importance on semantically meaningful words over frequently occurring ones in equation (19).

$$\text{CIDEr} = \frac{1}{K}\sum_{z=1}^{k} G_z \,.\log\left(\frac{a_z}{b_z}\right) \quad (19)$$

The part $\frac{1}{K}\sum_{z=1}^{k}$,take each word index Z (from **1 up to** K, where K is the number of reference captions or terms considered).Where (a_z) is the Count of the word (z)in the generated caption, the weighting factor (G_z), the number of reference captions $(k), and\ the$ count of word z in reference captions. If (a_z) is close to (b_z) then the ratio $\frac{a_z}{b_z}$is close to 1, and the log term is small → meaning the model matches human references well.

CIDEr is crucial for multimodal content generation, ensuring that SPO-IntelliConvNet creates video descriptions that match human-written captions based on user interactions. Since it penalizes generic descriptions and favours user-specific, personalized content. Figure 5 depicts the CIDEr performance. Figure 5 represents the CIDEr outcome.

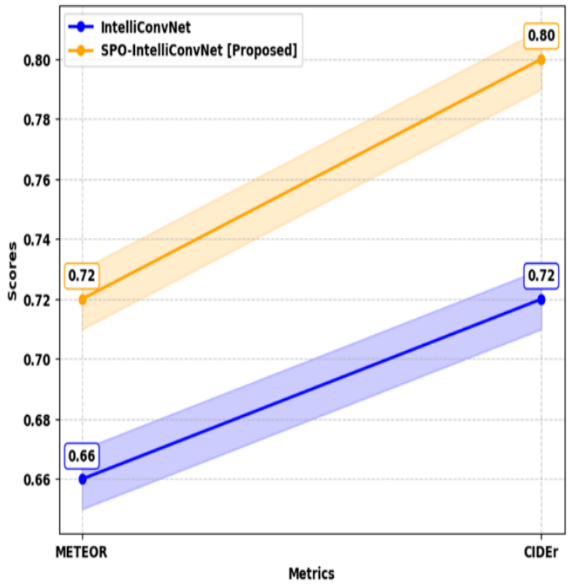

Figure 5: Visual representation of METEOR and CIDEr

In multimodal short video generation research, the traditional IntelliConvNet method has a CIDEr score of 0.72, while the proposed SPO-IntelliConvNet has a CIDEr score of 0.80, demonstrating a strong alignment between generated content and user-preferred references. This high score indicates that the SPO-IntelliConvNet model effectively captures contextual relevance, ensuring video content is personalized to user interactions. The gradual improvement in CIDEr across training epochs reflects enhanced semantic richness, proving the model generates engaging, high-quality video content tailored to user preferences.

4.4 Recall-oriented understudy for gisting evaluation ($ROUGE_L$)

Assesses recall-based similarity, confirming structural coherence and coverage of personalized video narratives compared to user-driven references. $ROUGE_L$ measures textual coherence by analyzing the longest common subsequence (LCS) between the reference text(N) and the generated (M). It emphasizes retrieval, ensuring that generated text retains essential context and fluency, making it ideal for evaluating text relevance and personalization in video descriptions, and mathematical representation in equation (20).

$$\mathrm{ROUGE_L} = \frac{LCS\,(M,N)}{length\,(M)} \tag{20}$$

$ROUGE_L$ evaluates the fluency and coherence of generated short video descriptions. A high $ROUGE_L$ score indicates that the proposed output preserves the structure and meaning of user-intended captions, improving video engagement. Figure 6 depicts the $ROUGE_L$ outcome.

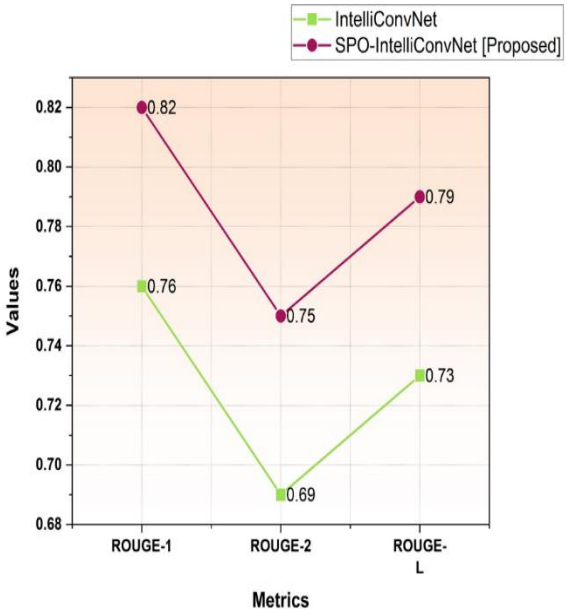

Figure 6: Visual representation of $ROUGE_L$

The proposed SPO-IntelliConvNet has a ROUGE-1 score (0.82), indicating strong word-level recall, ensuring key terms are retained in generated content. ROUGE-2 (0.75) highlights effective phrase-level coherence, while $ROUGE_L$ (0.79) confirms structural alignment. However, the traditional IntelliConvNet method has the ROUGE-1, ROUGE-2, and $ROUGE_L$ scores of 0.76, 0.69, and 0.73, respectively. These results demonstrate that SPO-IntelliConvNet successfully preserves meaning, fluency, and readability, enhancing personalized short video content generation with high textual relevance.

4.5 Training time

The training time quantifies the amount of time needed to optimize model parameters, indicating practicality, scalability, and computational performance for creating personalized content in real-time on adaptive short video platforms. The proposed SPO-IntelliConvNet method has a training time (7.0 hrs), whereas the conventional IntelliConvNet approach has a low training time (8.4), as shown in Table 3.

4.6 Inference speed

Inference speed quantifies the rate at which the model generates personalized outputs from user inputs, which has a direct impact on adaptability, user experience, and real-time response while delivering captivating video content. In comparison, the conventional IntelliConvNet strategy has an inference speed of 45 ms/frame, while the suggested SPO-IntelliConvNet has an inference speed of 36 ms/frame, as displayed in Table 3.

4.7 Resource consumption

The memory usage and CPU utilization metrics were used in the resource consumption evaluation. Memory usage evaluates the amount of computing power used for model training and inference, taking into account storage effectiveness, scalability, and optimization for managing multimodal data and real-time customization needs. With the low memory usage value of 7.3, the proposed SPO-IntelliConvNet outperformed the traditional IntelliConvNet method, which has a memory usage value of 8.2, as depicted in Table 3.

The proportion of processor resources used for content display, feature extraction, and model estimation is known as CPU utilization, and it represents workload balance, computational effectiveness, and system adaptability in real-time operation. When compared to the conventional approach, the proposed SPO-IntelliConvNet approach has a CPU utilization of 72%, while the conventional IntelliConvNet method has a CPU utilization of 77%, as shown in Table 3.

Table 3: Performance evaluation of training time, inference speed, memory usage, and CPU utilization

Methods	Training time (hrs)	Inference speed (ms/frame)	Memory usage (GB)	CPU utilization (%)
IntelliConvNet	8.4	45	8.2	77
SPO-IntelliConvNet [Proposed]	7.0	36	7.3	72

4.8 Discussion

The research concentrated on developing a multimodal short video generation model that personalizes content by analyzing user interactions, such as video views and comments. By integrating text, image, and audio modalities, the SPO-IntelliConvNet enhances content relevance and engagement. However, the research faces challenges like data dependency, computational complexity, contextual understanding, and scalability. These drawbacks are addressed by incorporating intelligent data filtering, adaptive optimization techniques, DL-based contextual learning, and a scalable architecture. The SPO-IntelliConvNet method integrated approach surpasses traditional models by enhancing multimodal feature extraction and real-time adaptation, ensuring improved content personalization and user engagement. The experiments were carried out in live content-generating situations to measure system delay and evaluate real-time personalization. The suggested SPO-IntelliConvNet outperformed traditional models in terms of user preference adaptation, exhibiting continuously reduced response times. This effectiveness can be achieved by the SPO-driven optimization of multimodal characteristics and the dimensionality reduction through PCA. The results demonstrate that the system can generate personalized short videos in real time without sacrificing user experience or quality. Current model evaluation is restricted to limited datasets, potentially affecting its generalizability across diverse short video domains. Future work will integrate broader, domain-specific datasets to enhance robustness and assess performance in personalized multimodal video generation.

5 Conclusion

The research introduces a new multi-modal content generation framework for personalized short video generation, which explains the present system restrictions of customization and real-time production. The model successfully combines contextual-relevant video content with user interactions by incorporating text, image and audio modalities using the SPO-IntelliConvNet. The method involves sophisticated pre-processing, which includes tokenizing of text, Canny edge detection of images, and Wiener filtering of audio, and then dimensionality reduction using PCA whilst maintaining important features. The proposed SPO-IntelliConvNet achieves superior performance by optimizing multimodal feature extraction, ensuring enhanced content personalization, semantic relevance, and linguistic accuracy in short video generation. Evaluation metrics, including BLEU, $ROUGE_L$, METEOR, and CIDEr, demonstrate the model's superior performance in generating personalized content. In comparison, the proposed SPO-IntelliConvNet method achieved the METEOR (0.72) and BLEU-4 (0.55), while the conventional IntelliConvNet approach achieved the METEOR (0.66) and BLEU-4 (0.49). These scores indicate strong textual coherence, semantic relevance, and effective personalization in short video content generation. The results indicate increased user satisfaction and engagement, highlighting the system's capability to enhance the short video experience through precise customization based on multimodal data.

5.1 Limitations and future scope

Limitations include computational complexity, high data dependency, and challenges in understanding deep contextual semantics, which can be addressed through advanced AI techniques and scalable architectures. Future improvements include enhancing real-time processing, integrating reinforcement learning for adaptive personalization, and expanding multimodal datasets for richer content generation.

Acknowledgements

This paper is an outcome of the 2023 Chongqing Municipal Education Science "14th Five-Year Plan" General Project "Construction and Reform Research on Digital Curriculum Resources for Media Arts Majors from the Perspective of Interdisciplinary Collaboration" (Project No.: K23YG2190393).

References

[1] Tang, W., & Zhang, S. (2025). Exploring the integration of augmented reality interfaces with short video content creation to enhance user interaction and engagement. *Journal of Computational Methods in Sciences and Engineering*, *25*(3), 2100-2111. https://doi.org/10.1177/14727978241309546

[2] Li, Min & Guo, Shujuan & Liu, Runchen. (2025). BERT-based Consumer Sentiment Analysis for Personalized Marketing Strategies. *Informatica*, 49. https://doi.org/10.31449/inf.v49i28.8232

[3] Mou, N., Jiang, Q., Zhang, L., Niu, J., Zheng, Y., Wang, Y., & Yang, T. (2022). Personalized tourist route recommendation model with a trajectory understanding via neural networks. *International Journal of Digital Earth*, *15*(1), 1738-1759. https://doi.org/10.1080/17538947.2022.2130456

[4] Wang, H. C., Maslim, M., & Hong, W. T. (2024). Personalized time-sync comment generation based on a multimodal transformer. *Multimedia Systems*, *30*(2), 105. https://doi.org/10.1007/s00530-024-01301-3

[5] Umale-Nagmote, A., Goel, C., & Lal, N. (2025). Enhanced Intelligent Video Monitoring using Hybrid Integration of Spatiotemporal Autoencoders and Convolutional LSTMs. *Informatica*, *49*(18). https://doi.org/10.31449/inf.v49i18.7502

[6] Wu, Y., Fu, R., Xing, T., Yu, Z. and Yin, F., 2025. A user behavior-aware multi-task learning model for enhanced short video recommendation. *Neurocomputing*, *617*, p.129076. https://doi.org/10.1016/j.neucom.2024.129076

[7] Li, H., Lin, J., Wang, T., Zhang, L., & Wang, P. (2022). A personalized short video recommendation method based on multimodal feature fusion. https://doi.org/10.21203/rs.3.rs-2033641/v1

[8] Qi, M. (2024). The short video platform recommendation mechanism based on the improved neural network algorithm to the mainstream media. *Systems and Soft Computing*, *6*, 200171. https://doi.org/10.1016/j.sasc.2024.200171

[9] Liu, Y., Xu, Y., & Liu, Z. (2024). Knowledge graph: a recommendation method for new media short videos. *Advances in Education, Humanities and Social Science Research*, *12*(1), 87-87. https://doi.org/10.56028/aehssr.12.1.87.2024

[10] Zhu, D. (2025). Optimizing the user personalized recommendation system of new media short videos by using machine learning. *Journal of Computational Methods in Sciences and Engineering*, 14727978251341490. https://doi.org/10.1177/14727978251341490

[11] Yang, Y. (2025). Short video content and recommendation algorithm based on deep learning. *Journal of Computational Methods in Sciences and Engineering*, 14727978251337878. https://doi.org/10.1177/14727978251337878

[12] Lu, Z., & Nam, I. (2021). Research on the influence of new media technology on internet short video content production under artificial intelligence background. *Complexity*, *2021*(1), 8875700. https://doi.org/10.1155/2021/8875700

[13] Song, J., & Liu, J. (2024). A data-driven short video international communication model based on indicator system communication network and attention BiLSTM neural network. *Scientific Reports*, *14*(1), 14532. https://doi.org/10.1038/s41598-024-65098-x