Received: 06 June 2026; **Revised:** 01 September 2026; **Accepted:** 07 September 2026; **Published:** 09 September 2026
Archs Sci. (2026) Volume 76(2), Pages 31-39, Paper ID 2026204.

The Application of New Communication and Video-Text Description Technologies to Advancing Privacy Protection in the Revitalization of Rural Cultural Industries

Liying Wang[1,*] and Jian Dong[2]
[1]School of Comics, Jilin Animation Institute, Changchun 130012, China
[2]Ani-Com School, Jilin University of the Arts, Changchun 130012, China
Corresponding author: Liying Wang (e-mail: sai89757@126.com).

Abstract This paper first examines the characteristics of social networks and the threats associated with privacy protection. Accordingly, the use of new communication technologies to construct a multifunctional heterogeneous network represents a potential means of addressing these challenges. This paper also investigates long-video description technology. First, considering the current lack of a dataset that satisfies the requirements of long-video description tasks, this paper designs and implements a complete set of processing methods for constructing a long-video description dataset according to the requirements of real-world scenarios. Chinese and English long-video description datasets are subsequently constructed using this method. In addressing the difficulties encountered during this process, it is necessary to recognize the value of rural culture and respond to people's needs, particularly their growing cultural needs. Attention should be given to practical implementation, local realities, and the appropriate direction of development. It is also necessary to overcome existing difficulties, stimulate cultural vitality, and provide a sustainable source of support for rural revitalization. The experimental results demonstrate that the optimized differential privacy algorithm contributes to privacy protection, supports the effective implementation of the proposed strategy, and facilitates the achievement of rural cultural development objectives.

Index Terms new communication technologies, video-text description technology, rural revitalization, privacy protection, optimized differential privacy algorithm

I. Introduction

Throughout the history of the Chinese nation, rural culture has remained rooted in agricultural land and has continuously evolved with the development of society. Rural culture continues to innovate, thereby contributing to the comprehensive revitalization of the countryside [1].

New wireless communication technologies, such as non-orthogonal multiple access, unmanned aerial vehicle (UAV) communications, energy harvesting, and free-space optical communications, have theoretically been demonstrated to effectively address the principal problems affecting current wireless networks. These technologies can also complement existing heterogeneous networks through cooperative diversity techniques, thereby facilitating the construction of more efficient, reliable, secure, and energy-efficient wireless communication systems [2]. As flexible communication platforms, UAVs can provide reliable transmission links in remote areas or disaster-affected regions with poor communication coverage and can help maintain uninterrupted communication during emergencies. At present, the common applications of UAVs in wireless communications can be divided into three categories: aerial base stations, aerial relays, and aerial information-transmission and data-collection systems. A video-text description algorithm aims to generate an accurate description of the content contained in given video data. The generated description should not only reflect the content of the video but should also be accurate and readable.

Image-based text-description tasks have achieved significant breakthroughs and have been applied in real-world settings. However, research on video-based text-description tasks remains challenging because, compared with a single image, video data contain more extensive information about scenes, actions, and temporal relationships. The principal approaches to video-text description include language-template-based methods, retrieval-based methods, and encoder–decoder-model-based methods [3]. At present, encoder–decoder-based video-text description algorithms constitute the mainstream approach to video-text description tasks. These algorithms generally consist of two principal components: an encoder and a decoder. The encoding component primarily uses a convolutional neural network to extract visual features from the input video data. The decoding component employs a recurrent neural network to decode the extracted visual features, generate word representations, and subsequently produce a

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

sequence of words. In this manner, a textual statement describing the corresponding video is generated [4]. Therefore, the research presented in this paper is conducted within an encoder–decoder modeling framework.

The first aspect of the research significance of rural cultural construction concerns its contribution to enriching theories of cultural development [5]. At present, most relevant studies in China discuss rural cultural development from a macro-level perspective. Since the beginning of the new era, the development of rural culture has received continuous support through national policies. Nevertheless, the development of rural culture in China has been constrained and affected by numerous factors. With appropriate policy support, rural cultural development can play an important role in rural revitalization. Practical experience can provide an empirical basis for scholars studying rural culture, enrich and develop theories of cultural construction, and enable these theories to provide more effective guidance for practice.

The study of rural culture should begin by examining its direction of development, aligning it with the requirements of the new era, conducting an in-depth investigation, and exploring an optimal development path. Such work is conducive to enriching and diversifying rural culture, providing a more substantial theoretical foundation, and further developing theories of cultural construction [6]. Rural cultural construction builds upon the achievements made in poverty alleviation. From a cultural perspective, identifying the distinctive value of rural culture and strengthening its role in rural social development can improve the cultural literacy and spiritual well-being of rural residents, contribute to the formation of civilized rural communities, and demonstrate the soft power of culture [7].

However, the current development of rural cultural construction does not always correspond to the level of rural economic development. Existing problems include inadequate infrastructure, insufficient participation by key stakeholders, limited vitality within rural cultural industries, and comparatively underdeveloped educational conditions [8]. It is therefore necessary to strengthen cultural development, obtain a deeper understanding of the principles of socialist modernization with Chinese characteristics and the distinctive characteristics of China's rural modernization, respond to farmers' aspirations for a better life, and satisfy the public's increasing spiritual and cultural needs.

In summary, the process of rural cultural construction requires recognition not only of the positive effects of rural cultural revitalization on rural development, the improvement of social customs, and the education of rural residents, but also of its relationship with other dimensions of rural development. It is necessary to systematically examine the problems arising during rural cultural construction in China and analyze their underlying causes. The revitalization of rural culture is not only an essential component of comprehensive rural revitalization but also an inherent requirement for strengthening the development of socialist culture with Chinese characteristics.

II. Optimization of the Differential Privacy Algorithm

When the length T of the finite data stream is 1 and $k = 1$, the data X_k can be directly protected using ϵ-differential privacy. Subsequently, Z_k is assigned to R_k and submitted to a trusted third party for data publication. The privacy budget at this point is $\epsilon_k = \epsilon$. This basic privacy protection algorithm is referred to as the Laplace perturbation algorithm (LPA). When $T > 1$, as time k advances, the LPA provides ϵ-differential privacy protection for a data stream with a finite length of T, and the privacy budget is $\epsilon_k = \epsilon/T$, where $k = 1, 2, \ldots, T$. As time progresses, the privacy budget allocated to each time point gradually decreases, the amount of noise introduced gradually increases, and the data utility of the LPA gradually declines. Because the Kalman filtering algorithm uses a recursive form in its calculations, it can process nonstationary random processes and thereby improve data utility. Considering the state-space model, the prediction process of the KFDP algorithm is presented in Eqs. (1) and (2):

$$\hat{\mathbf{x}}_k^- = \hat{\mathbf{x}}_{k-1}, \tag{1}$$

$$\mathbf{P}_k^- = \mathbf{P}_{k-1} + \mathbf{Q}. \tag{2}$$

The correction process of the KFDP algorithm is presented in Eqs. (3)–(6):

$$\mathbf{z}_k = \mathbf{x}_k + \mathbf{v}_k, \tag{3}$$

$$\mathbf{K}_k = \mathbf{P}_k^- \left(\mathbf{P}_k^- + \mathbf{R}\right)^{-1}, \tag{4}$$

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k \left(\mathbf{z}_k - \hat{\mathbf{x}}_k^-\right), \tag{5}$$

$$\mathbf{P}_k = \left(\mathbf{I} - \mathbf{K}_k\right) \mathbf{P}_k^-, \tag{6}$$

The calculation procedure is presented in Eqs. (7) and (8):

$$\chi_{k-1}^{(i)} = \begin{cases} \hat{\mathbf{x}}_{k-1}, & i = 0, \\ \hat{\mathbf{x}}_{k-1} + \left(\sqrt{(n+\lambda)\mathbf{P}_{k-1}}\right)_i, & i = 1, 2, \ldots, n, \\ \hat{\mathbf{x}}_{k-1} - \left(\sqrt{(n+\lambda)\mathbf{P}_{k-1}}\right)_i, & i = n+1, n+2, \ldots, 2n, \end{cases} \tag{7}$$

$$\begin{cases} \omega_m^{(0)} = \dfrac{\lambda}{n+\lambda}, \\ \omega_c^{(0)} = \dfrac{\lambda}{n+\lambda} + \left(1 - \alpha^2 + \beta\right), \\ \omega_m^{(i)} = \omega_c^{(i)} = \dfrac{\lambda}{2(n+\lambda)}, \qquad i = 1, 2, \ldots, 2n. \end{cases} \tag{8}$$

The calculation procedure is similar to that of a standard LSTM, and the corresponding formulas are presented in Eqs. (9)–(14):

$$i_t = \sigma\left(F'_{it} + h'_{it-1} + X'_i + b_i\right), \tag{9}$$

$$f_t = \sigma\left(F'_{ft} + h'_{ft-1} + X'_f + b_f\right), \tag{10}$$

$$o_t = \sigma\left(F'_{ot} + h'_{ot-1} + X'_o + b_o\right), \tag{11}$$

$$c'_t = \mathrm{Tanh}\left(F'_{ct} + h'_{ct-1} + X'_c + b_c\right), \tag{12}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot c'_t, \tag{13}$$

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

$$h_t = o_t \cdot \mathrm{Tanh}\left(c_t\right). \tag{14}$$

The prediction process of the UKFDP algorithm is presented in Eqs. (15)–(20):

$$\begin{aligned}\chi_{k-1}^{(i)} = \Bigg[&\hat{\mathbf{x}}_{k-1}, \hat{\mathbf{x}}_{k-1} + \left(\sqrt{(n+\lambda)\mathbf{P}_{k-1}}\right)_1, \ldots, \\ &\hat{\mathbf{x}}_{k-1} + \left(\sqrt{(n+\lambda)\mathbf{P}_{k-1}}\right)_n, \\ &\hat{\mathbf{x}}_{k-1} - \left(\sqrt{(n+\lambda)\mathbf{P}_{k-1}}\right)_{n+1}, \ldots, \\ &\hat{\mathbf{x}}_{k-1} - \left(\sqrt{(n+\lambda)\mathbf{P}_{k-1}}\right)_{2n}\Bigg], \ \chi_k^{(i)} = \chi_{k-1}^{(i)}, \end{aligned} \tag{15}$$

$$\mathbf{X}_k^{(i)} = \mathbf{X}_{k-1}^{(i)}, \tag{16}$$

$$\hat{\mathbf{x}}_k^- = \sum_{i=0}^{2n} \omega_m^{(i)} \chi_k^{(i)}, \tag{17}$$

$$\mathbf{P}_k^- = \sum_{i=0}^{2n} \omega_c^{(i)} \left[\chi_k^{(i)} - \hat{\mathbf{x}}_k^-\right]\left[\chi_k^{(i)} - \hat{\mathbf{x}}_k^-\right]^{\mathrm{T}} + \mathbf{Q}, \tag{18}$$

$$\mathbf{z}_k = \chi_k^{(i)} + \mathbf{v}_k, \tag{19}$$

$$\hat{\mathbf{z}}_k^- = \sum_{i=0}^{2n} \omega_m^{(i)} \mathbf{z}_k. \tag{20}$$

III. Methods

In the video-text description task, the visual content contained in a video is converted into descriptive textual statements as output. In recent years, in research related to video-text description, some researchers have proposed using visual semantic detectors to help decoding models generate more accurate descriptions during the decoding process. Accurate semantic information can improve the quality of the generated textual descriptions, whereas inaccurate semantic information can cause the network to incorrectly represent the video content. The specific structure is illustrated in Figure 1.

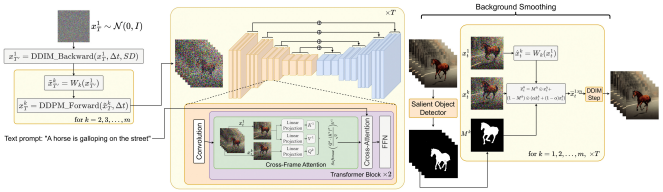

Figure 1: Optimization Diagram of the Video-Text Description Structure

Therefore, the focus of this paper primarily includes the following two aspects. The term NEEE in Figure 2 represents the textual input.

This study explores effective approaches and methods for rural cultural construction in China and provides corresponding suggestions for its advancement. Specifically, this paper proposes countermeasures and recommendations covering four major aspects. Rural cultural construction is a long-term and systematic undertaking that requires the government to attach considerable importance to the process and fully

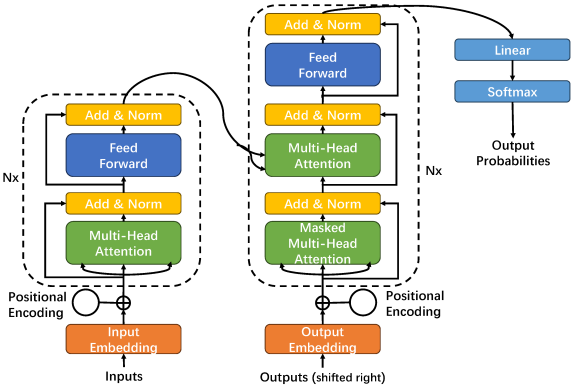

Figure 2: Optimization of the BERT Model

mobilize all relevant positive factors. Conducting research at the theoretical level is also a demanding undertaking. It is necessary to organize relevant data, conduct field research, identify existing problems, and determine the most appropriate and effective pathway for rural cultural construction.

Figure 3 presents the framework of the mainstream encoder–decoder model for video-text description. First, the input video-frame sequence, containing sT frames, is encoded by the convolutional neural network (CNN) encoder to obtain the visual video feature X.

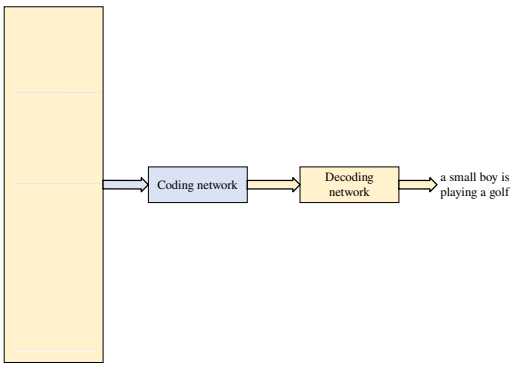

Figure 3: Optimized Video-Text Description Framework Based on an Encoder–Decoder Model

Because video-text description is a challenging task, early video-text description algorithms primarily employed language-template-based methods, which generated fixed sentence templates to describe video content and consequently produced descriptions with limited sentence patterns. Because the task involves two modalities, namely text and video, the most commonly used method in this field is based on the encoder–decoder network framework. Through the encoder, the different modalities of text and video are mapped into the same feature space, after which text representing the video content is generated through the decoding process. Figure 4 presents the video-text description framework of the semantic-feature-based encoder–decoder model. Within this framework, the encoding network is first used to encode the

input video sequence and obtain its visual features. These visual features are then entered into the semantic encoding network to obtain the high-level semantic features of the video. Finally, the semantic features are embedded in the decoding network to assist its operation, thereby improving the decoding performance of the model and making the generated textual descriptions more accurate.

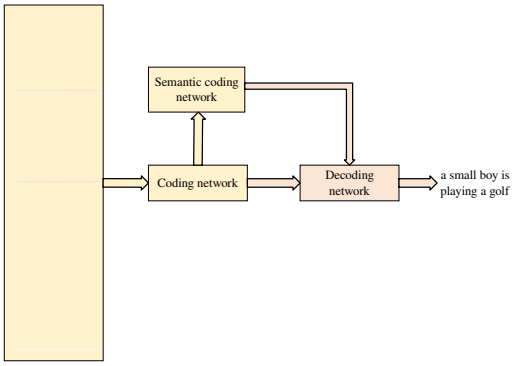

Figure 4: Optimized Framework of an Encoder–Decoder Model Based on Semantic Features

The channel is an important component of a communication system. A fixed communication channel is referred to as a constant-parameter channel because it maintains the same transmission characteristics over a prolonged period. Its transmission media include twisted-pair cables, coaxial cables, and optical fibers. The principal difference between wireless and fixed communication is that the transmission path between the transmitter and receiver in wireless communication constantly changes. Regardless of the type of propagation path, the propagation characteristics of wireless communication exhibit considerable randomness; therefore, this type of channel is referred to as a random channel. Social-network data are characterized by high dimensionality and substantial complexity. Therefore, to make privacy-protection technology applicable to more complex, large-scale social-network systems, research on privacy-protection technologies must consider the requirements of data publication. Such algorithms must resolve the conflict between protecting the privacy of user attributes and relationships and minimizing the corresponding loss of information.

The enthusiasm and initiative of the public are essential driving forces for ensuring effective cultural construction. For a long time, China has adhered to the principle of promoting socialist core values as the standard for public communication and guidance and has made considerable efforts to publicize the current principles and policies governing the Party's work in rural areas. Various approaches have been adopted across different areas of rural cultural construction. Regarding communication channels, radio and television broadcasts, billboards displayed in outdoor public spaces, cultural publicity walls, and other methods have been used for promotional purposes. Through various forms of thematic education, including literary and artistic performances and rural bookstores, education concerning love for the Party, patriotism, and commitment to the family has been integrated into the daily lives of the general public. These activities have made important contributions to shaping values related to moral conduct, integrity, and filial responsibility within the family.

In summary, the rich traditional values embodied in rural culture constitute its core components and reflect the long cultural history and profound cultural foundations of the Chinese nation. On the basis of rural civilization, excellent cultural achievements from other societies should be appropriately incorporated. Priority should be given to protection and inheritance, supplemented by innovation and development, while preserving valuable elements and eliminating inappropriate or outdated elements. The development of traditional rural culture should be continuously refined. Rural cultural construction promotes and advances socialist core values, ensures substantive progress in building a civilized rural ethos in the new era, and contributes to realizing the strategic blueprint for rural revitalization.

IV. Experiments

A. Experimental Objectives

The primary objective of this experiment is to validate the effectiveness of the three proposed video semantic-feature enhancement models (MLP-I, MLP-II, and MLP-III) in the video-caption generation task. Specifically, by comparing the effects of semantic features generated using different enhancement strategies on the performance of downstream decoder models across the standard benchmark datasets MSVD and MSR-VTT, we aim to assess the contribution of each semantic-enhancement strategy to the quality of the generated textual descriptions. In addition, we evaluate the proposed differential-privacy optimization algorithms, including KFDP and DPRKF, in terms of their ability to preserve data utility under varying privacy-budget conditions. This experimental design therefore evaluates both the quality of video-caption generation and the utility-preservation performance of the investigated privacy-protection algorithms.

B. Datasets

The experiments were conducted using two widely used benchmark datasets:

MSVD (Microsoft Video Description). This dataset contains approximately 2,000 short video clips, each of which is annotated with multiple textual descriptions written by human annotators. The video content typically covers scenes from everyday life, and the accompanying sentences are generally short and semantically clear. These characteristics make MSVD a standard benchmark dataset for evaluating the performance of video-captioning methods.

MSR-VTT (Microsoft Research Video to Text). This dataset includes more than 10,000 video clips and approximately 200,000 descriptive sentences. Its video content is more diverse and complex than that of MSVD and covers a broad range of semantic categories. The MSR-VTT dataset is therefore used to assess the generalization ability of the proposed

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

models in large-scale and comparatively challenging video-captioning scenarios.

Figure 5 presents the average relative errors obtained by the evaluated algorithms under different privacy budgets. The numerical values corresponding to this figure are also reported in Table 1.

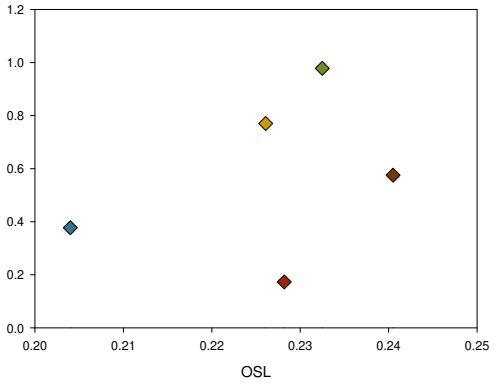

Figure 5: Average Relative Error of Each Algorithm Under Different Privacy Budgets

Figure 5 and Table 1 present the average relative errors of the evaluated algorithms. The results indicate that each algorithm provides greater data utility under a lower level of privacy protection. When $\epsilon = 0.1$, the average relative errors of the DPRKF, η-EDPRF, and θ-HDPRF algorithms are 2.8611, 2.5135, and 2.4948, respectively. When $\epsilon = 0.9$, the average relative errors of the DPRKF, η-EDPRF, and θ-HDPRF algorithms decrease to 0.9419, 0.8287, and 0.8287, respectively. On the other datasets, the data utility provided by each algorithm follows the same general trend. The simulation results therefore indicate that data utility increases as the level of privacy protection decreases.

Table 1: Average Relative Error of Each Algorithm Under Different Privacy Budgets

Dataset	ϵ	LPA	KFDP	BKFDP	IKFDP	DPRKF	η-EDPRF	θ-HDPRF
OS1	0.2	16.107	15.540	14.667	14.034	2.861	2.513	2.494
	0.4	4.508	4.199	4.173	4.141	1.754	1.547	1.556
	0.6	3.591	3.339	3.239	3.216	1.222	1.201	1.191
	0.8	2.136	1.975	1.973	1.971	0.977	0.839	0.823
	0.9	1.722	1.616	1.611	1.606	0.941	0.828	0.828
PS2	0.2	5.636	5.328	5.143	5.300	1.487	1.228	0.212
	0.4	1.826	1.707	1.698	1.709	1.056	1.027	1.022
	0.6	1.158	1.095	1.082	1.049	0.724	0.730	0.723
	0.8	0.835	0.771	0.767	0.744	0.644	0.630	0.628
	0.9	0.614	0.565	0.567	0.608	0.565	0.565	0.565
OS3	0.2	2.791	2.616	2.565	2.575	0.198	0.189	0.191
	0.4	1.059	1.037	0.981	0.982	0.185	0.189	0.189
	0.6	0.515	0.518	0.474	0.479	0.201	0.189	0.190
	0.8	0.379	0.426	0.347	0.342	0.188	0.187	0.189
	0.9	0.342	0.319	0.317	0.314	0.196	0.189	0.188
OS4	0.2	3.025	2.860	0.772	2.751	0.225	0.214	0.214
	0.4	0.994	0.918	0.913	0.926	0.220	0.214	0.214
	0.6	0.619	0.547	0.544	0.562	0.212	0.214	0.213
	0.8	0.417	0.376	0.346	0.361	0.215	0.214	0.214
	0.9	0.371	0.340	0.317	0.327	0.220	0.214	0.215

C. Model Setup and Comparison Methods

This experiment compares three encoder models developed through different enhancements to the Semantic Detection Network (SDN):

MLP-I Model. This model incorporates a Highway-layer structure into the SDN to enhance the non-linear representational capacity of the extracted semantic features.

MLP-II Model. This model introduces a semantic-word difference-enhancement module into the SDN to emphasize important disparities among semantic features.

MLP-III Model. This model combines the enhancements introduced in both MLP-I and MLP-II by integrating the Highway layer and the semantic-difference module, thereby comprehensively improving the quality of semantic representations.

During the training process, the models are saved at different iteration stages. The semantic features generated by these saved models are subsequently passed to a fixed decoder network. This procedure enables the effectiveness of the different semantic features in downstream text generation to be evaluated under uniform decoding conditions.

D. Evaluation Metrics

The quality of the generated semantic features is assessed using the mean average precision (MAP) metric, which reflects the contribution of the semantic vectors to the performance of the description-generation task. A higher MAP value indicates that the generated semantic features provide more accurate information for the downstream decoder and contribute more effectively to the generation of appropriate textual descriptions.

The performance of the differential-privacy algorithms is measured using the average relative error. This metric indicates how the utility of the protected data changes under different privacy-budget levels, represented by ϵ. A lower average relative error reflects the stronger preservation of data utility after the application of the relevant differential-privacy algorithm.

E. Experimental Results and Analysis

Table 2 presents the MAP values obtained by the different models at various training iterations on the MSVD dataset.

Table 2: MAP Values Obtained on the MSVD Dataset

Dataset	Method	100	200	400	800	1000
MSVD	MLP-I	0.2763	0.3958	0.5467	0.6431	0.6607
	MLP-II	0.2638	0.3428	0.4344	0.5124	0.5389
	MLP-III	0.2825	0.3853	0.5259	0.6478	0.6783

As shown in Table 2 and illustrated in Figure 6A, the MLP-III model generally achieves stronger performance than the other two models across the reported training stages and attains the highest MAP score of 0.6783 at the 1,000th iteration. This result indicates that the semantic features generated by MLP-III can effectively enhance the quality of the generated video descriptions. Although the relative performance varies during the earlier training iterations, the final result obtained

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

by MLP-III demonstrates the advantage of integrating the two proposed semantic-feature enhancement mechanisms.

Table 3 presents the MAP values obtained by the different models at various training iterations on the MSR-VTT dataset.

Table 3: MAP Values Obtained on the MSR-VTT Dataset

Dataset	Method	100	200	400	800	1000
MSR-VTT	MLP-I	0.1987	0.3104	0.4272	0.5003	0.5331
	MLP-II	0.1825	0.2713	0.3898	0.4526	0.4814
	MLP-III	0.2146	0.3489	0.4765	0.5613	0.5849

The results obtained on the more complex MSR-VTT dataset, as presented in Table 3, further demonstrate the strong generalization ability of the MLP-III model in large-scale video-captioning tasks. In particular, MLP-III achieves the highest MAP value at every reported training iteration, increasing from 0.2146 at the 100th iteration to 0.5849 at the 1,000th iteration.

Table 4: Experimental Results of Different Models Generating Semantic Features on the MSVD Dataset

Dataset	Iterations MAP Method	100	200	400	800	1000
MSVD	MLP-I	0.2763	0.3958	0.5467	0.6431	0.6607
	MLP-II	0.2638	0.3428	0.4344	0.5124	0.5389
	MLP-III	0.2825	0.3853	0.5259	0.6478	0.6783

Table 5: Experimental Results of Different Models Generating Semantic Features on the MSR-VTT Dataset

Dataset	Iterations MAP Method	100	200	400	800	1000
MSR-VTT	MLP-I	0.1927	0.2248	0.2665	0.2889	0.2699
	MLP-II	0.1615	0.2191	0.2527	0.2907	0.2987
	MLP-III	0.1995	0.2327	0.2806	0.3236	0.3172

The experimental data reported in Tables 4 and 5 present the results obtained by training the three video semantic-feature enhancement encoder models proposed in this paper. During the continuous iterative training of the network, multiple intermediate models were saved, and these saved models were used to generate semantic features on the standard public MSVD and MSR-VTT datasets. The MAP values obtained from these generated semantic features are reported at several training iterations. The MLP-I model refers to the model in which a Highway-layer structure is added to the reference Semantic Detection Network used for semantic-feature extraction. The MLP-II model refers to the model in which a video semantic-word difference-amplification module is added to the reference Semantic Detection Network. The MLP-III model refers to the model in which both the Highway-layer structure and the video semantic-word difference-amplification module are added to the reference Semantic Detection Network. This combined structure is intended to improve semantic-feature extraction by exploiting the complementary functions of the two enhancement mechanisms.

Table 6 shows that, compared with the other two models proposed in this paper, the semantic features generated by the MLP-III model can more effectively assist the decoder model in improving the accuracy of the generated textual descriptions. Among the evaluation metrics used to assess the textual descriptions generated by the decoder model, CIDEr, METEOR, and the other reported evaluation measures reach their highest or most competitive values when the semantic features generated by the MLP-III model are used.

Table 6: Experimental Comparison Results on the MSVD Dataset

Dataset	Method	MAP	B-1	B-2	B-3	B-4	C	M	R
MSVD	SAM-SS	0.7415	–	–	–	62.5	109.8	39.1	77.1
	SAM-SS	0.4756	–	–	–	61.9	103.1	37.9	76.9
	Ours	0.6607	89.5	80.5	71.8	62.6	109.7	39.4	77.2
	Ours	0.5389	87.3	77.4	68.5	59.4	106.8	38.8	76.5
	Ours	0.6783	89.4	81.1	72.1	62.9	110.9	39.6	77.1

With the acceleration of urbanization, rural culture in many regions has exhibited increasing weakness, decline, and deterioration in traditional village customs. Monitoring and investigation reports concerning migrant workers over multiple years indicate that the number of workers migrating from rural areas continues to increase. Consequently, the phenomenon of hollow villages caused by the departure of young people has become increasingly prominent and represents a major challenge confronting current rural cultural construction. First, an increasingly evident and potentially irreversible pattern of population outflow has gradually developed in rural Chinese society, particularly in economically underdeveloped regions. Most of the people remaining in these communities are older adults and children, who cannot independently constitute the principal workforce required for rural cultural construction. Second, the hollowing out of rural communities involves not only population loss but also the weakening of the spiritual and cultural lives of rural residents. Moreover, owing to the long-term implementation of the urban–rural dual system in China, numerous economic, educational, technological, and cultural resources have tended to flow toward urban areas. These conditions have further weakened the human and material foundations required for sustainable rural cultural development.

Figure 6 presents comparisons of the training performance of two semantic encoder structures: EA-EM plastic and ES-EA-EM plastic. Subfigures (A) and (C) present the episode-wise step distributions for the best-performing seed, Seed-6, of each model. Individual points represent the raw numbers of steps recorded for the episodes, whereas the orange and green curves represent the moving median and moving mean, respectively, calculated over a window of 100 episodes. The EA-EM plastic structure exhibits greater variance and some improvement during the later stages of training, whereas the ES-EA-EM plastic structure demonstrates a comparatively stable and consistent training pattern. Subfigures (B) and (D) present the performance of 10 randomly seeded models for each architecture, with the results averaged over every 100 episodes. The EA-EM plastic structure exhibits greater peak variance and more instability across the different seeds, whereas the ES-EA-EM plastic structure demonstrates

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

smoother convergence and lower sensitivity to random initialization, indicating stronger generalization ability. Subfigure (E) compares the maximum averaged number of steps per episode achieved across the seeds. Although EA-EM plastic exhibits a slightly higher median value, its distribution is more dispersed. In contrast, ES-EA-EM plastic produces more consistent results; however, the statistical test indicates that the difference between the two structures is not significant (n.s.). Overall, ES-EA-EM plastic provides better training stability and robustness, while both models remain competitive in terms of their peak performance capabilities.

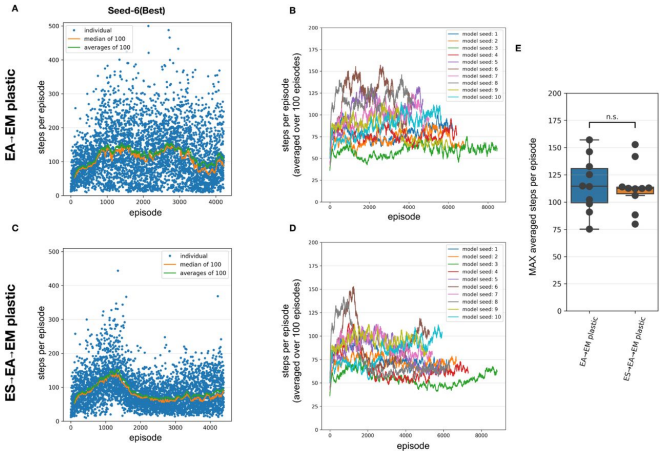

Figure 6: Training Performance Trends of Different Semantic Encoder Structures on the MSVD Dataset

Figure 7 illustrates the final evaluation results of the proposed model on two benchmark datasets—MSVD, represented by the green bars, and MSR-VTT, represented by the blue bars—using four widely adopted video-captioning evaluation metrics: BLEU@4, METEOR, CIDEr, and ROUGE.

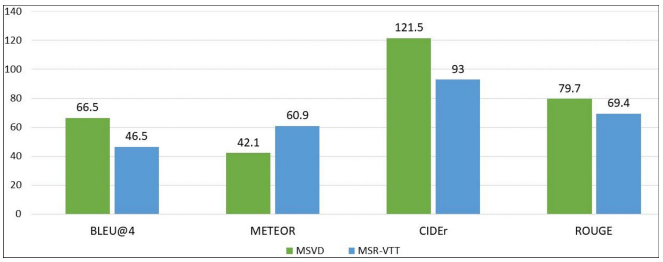

Figure 7: Final Performance Comparison on the MSVD and MSR-VTT Datasets Using Common Language Metrics

Across most of the reported metrics, the model demonstrates stronger performance on the MSVD dataset. It obtains a BLEU@4 score of 66.5 on MSVD, which is substantially higher than the score of 46.5 obtained on MSR-VTT. The same general pattern is evident for CIDEr: the model achieves a score of 121.5 on MSVD, compared with 93 on MSR-VTT, indicating stronger agreement with the corresponding human-generated annotations. For ROUGE, the model obtains a score of 79.7 on MSVD, whereas it achieves a score of 69.4 on MSR-VTT. An exception to this pattern is observed for the METEOR metric, for which MSR-VTT outperforms MSVD, with respective scores of 60.9 and 42.1. This result suggests stronger alignment on MSR-VTT in terms of synonym recognition, semantic correspondence, and paraphrase matching.

These results indicate that the model generally performs more favorably on the MSVD dataset, possibly because MSVD contains comparatively simpler video content and uses a more consistent annotation style. By contrast, the greater complexity and variability of the video and linguistic content in MSR-VTT make this dataset more challenging for the model. Nevertheless, the overall performance remains strong across both benchmark datasets, highlighting the robustness, adaptability, and generalization potential of the proposed model under different video-captioning conditions.

Figure 8 presents qualitative results comparing different video-to-text or video-transformation models under two distinct editing instructions: “color his dress white” and “make it Van Gogh Starry Night style”. The figure contains output sequences from four sources: the original video, Video Instruct-Pix2Pix (Ours), Instruct-Pix2Pix, and Tune-A-Video.

Figure 8: Video Captioning Output Comparison Across Methods (Qualitative Results)

The first row presents the original, unedited video frames showing a person performing a breakdance movement and a skier descending a snow-covered slope.

The second row presents the results obtained using the proposed Video Instruct-Pix2Pix method. For the instruction “color his dress white”, the proposed method demonstrates a high degree of conformity with the instruction by producing smooth and visually coherent white clothing across all of the displayed frames. Similarly, in the “Van Gogh Starry Night style” task, the proposed method not only transfers the requested artistic style but also effectively preserves the semantic layout, principal objects, and movement of the skier throughout the video sequence.

By comparison, the Instruct-Pix2Pix baseline produces generally appropriate edits but exhibits occasional inconsistencies in color and object-boundary details. Some frames appear oversimplified, and their temporal consistency is slightly weaker. The Tune-A-Video method applies a more aggressive form of style transfer; however, during this process, it distorts some object boundaries and disrupts the continuity of motion. These limitations are particularly visible in the artistic-style example, in which the ski trails and human figures become less

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

clearly distinguishable. The qualitative comparison therefore indicates that the proposed method provides a more favorable balance between instruction-based editing, preservation of semantic content, and temporal consistency across successive video frames.

Figure 9 presents an ablation study evaluating the effectiveness of two principal mechanisms in the video-generation framework: motion modeling in the latent space and cross-frame attention. Four model variants are compared across several tasks, including horse-motion synthesis, portrait generation, sketch-to-video generation, and pose-guided animation. In the first row, in which both motion modeling in the latent space and cross-frame attention are removed, the outputs exhibit severe temporal inconsistency, including shape jitter and content drift across consecutive frames. Introducing motion modeling in the latent space alone, as shown in the second row, improves temporal smoothness, particularly in dynamic scenes such as the galloping-horse example. However, fine details and the consistency of identity and structure remain unstable. The third row, which employs cross-frame attention without latent-space motion modeling, achieves stronger spatial alignment, as demonstrated by the improved preservation of poses and portraits, but it still lacks a coherent representation of motion. The final row, in which both motion-aware latent representations and cross-frame attention are enabled, produces the most temporally consistent and semantically faithful results. This configuration maintains structural integrity and smooth transitions across all of the presented examples. These findings demonstrate that the two components are both necessary and complementary and jointly contribute to the model's ability to generate high-quality, semantically appropriate, and temporally coherent video outputs.

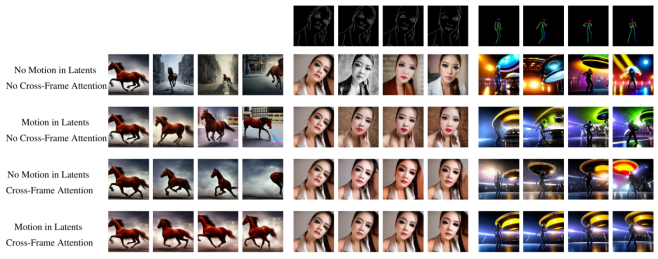

Figure 9: Ablation Study on Motion Latents and Cross-Frame Attention

F. Validation of Differential-Privacy Algorithm Utility

The average relative errors of several differential-privacy algorithms, including KFDP, IKFDP, and DPRKF, were evaluated under different privacy budgets, denoted by ϵ, across four datasets: OS1, OS2, OS3, and OS4. The comparative results are presented in Figure 10. These results are used to assess the extent to which each privacy-protection method can preserve the utility of the original data while providing the required level of privacy protection.

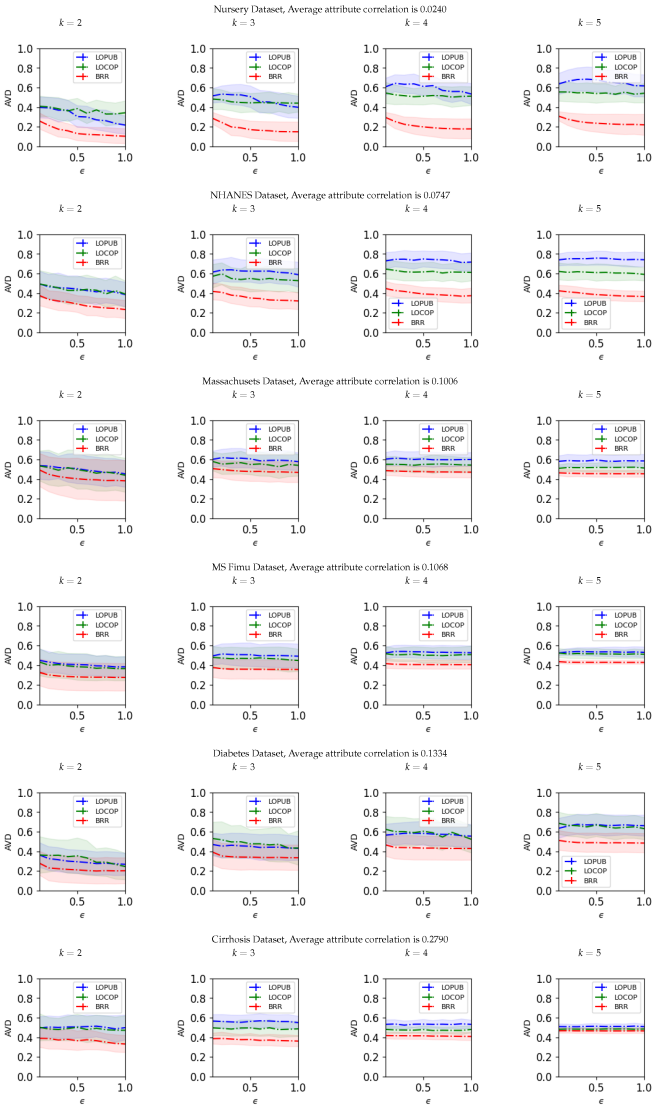

Figure 10: Comparison of Average Relative Errors Across Differential-Privacy Algorithms Under Varying Privacy Budgets (ϵ)

As summarized in Table 7, the DPRKF algorithm achieves an average relative error of 0.8287 when $\epsilon = 0.9$, demonstrating its strong ability to maintain high data utility under the evaluated privacy conditions. The results obtained at the two reported privacy-budget values also indicate that the average relative errors decrease as ϵ increases.

Table 7: Summary of Relative Errors Under Selected Privacy Budgets

ϵ Value	KFDP Error	DPRKF Error	HDPRF Error
0.1	2.8611	2.5135	2.4948
0.9	0.9419	0.8287	0.8287

Figure 10 also presents a comprehensive evaluation of the average relative error, or AVD, obtained by three differential-privacy algorithms—LOPUB, LOCOP, and BRR—across six

datasets: Nursery, NHANES, Massachusetts, MS Fimu, Diabetes, and Cirrhosis. The figure reports results under four clustering settings, namely $k = 2, 3, 4, 5$. The results demonstrate that BRR consistently obtains the lowest AVD across the reported datasets and ϵ values, indicating superior preservation of data utility under the evaluated privacy constraints. LOCOP demonstrates moderate performance, whereas LOPUB generally produces the highest AVD, particularly when ϵ is small. This result indicates a substantial reduction in data usability under more restrictive privacy budgets. As ϵ increases, all of the evaluated methods demonstrate improved utility because of the corresponding relaxation of the privacy constraints. Nevertheless, BRR converges more rapidly and remains comparatively stable across the different experimental scenarios.

Furthermore, the influence of attribute correlation is evident in the reported results. LOPUB exhibits greater deterioration in performance on datasets with relatively high levels of correlation, such as the Cirrhosis dataset, for which the reported correlation is 0.2790. By contrast, BRR maintains comparatively robust performance regardless of the reported correlation level. This analysis highlights the strong generalization ability of BRR and its effectiveness in preserving data utility across diverse privacy budgets, clustering configurations, dataset characteristics, and attribute-correlation conditions. Consequently, the reported results suggest that BRR constitutes a comparatively reliable option for privacy-preserving data publication under the experimental conditions considered in this study.

V. Conclusion

This paper focuses on NOMA, UAV communications, FSO communications, and other emerging 5G communication technologies and conducts relevant theoretical research by integrating key technologies such as cooperative communication, spatial point processes, physical-layer security, and energy harvesting. Video-description technology must address the complexity and diversity of video content, perform high-level semantic analysis of videos, and generate descriptive statements in natural language. Based on the long-video description model proposed above, this paper further investigates the role of visual information and proposes three types of visual-feature tags: video-category tags, object-detection tags, and key-person-detection tags. It also employs two methods to fuse visual tags with textual information, thereby enabling visual information to support textual information in completing the long-video description task. The experimental results demonstrate that integrating visual and textual information can improve the effectiveness and accuracy of long-video description generation.

Data Availability

The experimental data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding this work.

Funding Statement

This work was supported by the "13th Five-Year Plan" Social Science Project of the Education Department of Jilin Province, entitled "Research on the Construction Strategy for the 'Land Feature Changbai' Animation Creative Brand in Jilin Province under the Background of Rural Revitalization" (Project No. JJKH20201228SK).

This work was also supported by the project entitled "Research on the Practical Teaching Model of Animation Education for Empowering Rural Revitalization in Jilin," funded by the Jilin Higher Education Society (Project No. JGJX2022D262).

References

[1] Sun, Y., Liu, J., Wang, J., Cao, Y., & Kato, N. (2020). When machine learning meets privacy in 6G: A survey. IEEE Communications Surveys & Tutorials, 22(4), 2694–2724.

[2] Pencarelli, T. (2020). The digital revolution in the travel and tourism industry. Information Technology & Tourism, 22(3), 455–476.

[3] del Gaudio, G., Refugio, C. N., Jurcic, I., Della Corte, V., Ferdinand-James, D., Said, M. M. T., Sawicka, B., Mohan, T. R., Aravind, V. R., Umachandran, K., & Amuthalakshmi, P. (2019). Designing learning-skills towards Industry 4.0. World Journal on Educational Technology: Current Issues, 11(2), 150–161.

[4] Ji, B., Wang, Y., Song, K., Li, C., Wen, H., Menon, V. G., & Mumtaz, S. (2021). A survey of computational intelligence for 6G: Key technologies, applications and trends. IEEE Transactions on Industrial Informatics, 17(10), 7145–7154.

[5] Zhang, C., Roh, B.-H., & Shan, G. (2023). Poster: Dynamic clustered federated framework for multi-domain network anomaly detection. In Companion of the 19th International Conference on Emerging Networking EXperiments and Technologies (pp. 71–72). Association for Computing Machinery.

[6] Antonucci, F., Figorilli, S., Costa, C., Pallottino, F., Raso, L., & Menesatti, P. (2019). A review on blockchain applications in the agri-food sector. Journal of the Science of Food and Agriculture, 99(14), 6129–6138.

[7] Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29.

[8] Silvestri, S., Richard, M., Edward, B., Dharmesh, G., & Dannie, R. (2021). Going digital in agriculture: How radio and SMS can scale-up smallholder participation in legume-based sustainable agricultural intensification practices and technologies in Tanzania. International Journal of Agricultural Sustainability, 19(5–6), 583–594.