

Article

An Algorithm for Fine-Grained Content Extraction and Understanding in Short Videos

Yanqi Wan ^{1,†}, Shuya Zhang ^{2,†} , Yi Xu ³, Kunfang Zhang ⁴, Heyi Wang ⁵ and Mingzheng Liu ^{1,*} 

¹ School of Journalism and Communication, Jilin University, Changchun 130012, China; wanyq22@mails.jlu.edu.cn

² School of Journalism and Information Communication, Huazhong University of Science and Technology, Wuhan 430074, China; shuyazhang0821@163.com

³ School of Public Administration, Jilin University, Changchun 130012, China; 15041161345@163.com

⁴ School of Economics, Jilin University, Changchun 130012, China; zhangkf7567@126.com

⁵ School of Resources and Environment, Jilin Agricultural University, Changchun 130118, China; 16601399319@163.com

* Correspondence: liumz@jlu.edu.cn

† These authors contributed equally to this work.

Abstract

We integrated communication theory with advanced computer vision techniques to propose a novel approach for fine-grained content extraction from short videos. Unlike methods focused on summarization or subtitle generation for longer videos, our approach emphasizes extracting detailed content and understanding the intricate narrative structure of short videos. By employing scene segmentation, similarity-based filtering algorithms, and support vector machines, the method identifies keyframes that capture precise visual details. Further, it generates semantically accurate textual descriptions using the mPLUG model, enabling an in-depth understanding of video content. Using a dataset of short videos from the cultural and tourism domain, we validated the proposed method. Experimental results demonstrate that our approach achieves high precision in identifying and understanding detailed visual elements, effectively bridging the gap between visual representation and semantic meaning. Additionally, the study explores the influence of different video content types, interference factors, and image description models on fine-grained content extraction, highlighting its potential for improving intelligent analysis of short-video data.

Keywords: short videos; computer vision; fine-grained video captioning; keyframe extraction; multimodal learning; text generation



Academic Editors: Daqian Shi, Wei Kong and Yuqi Ouyang

Received: 18 May 2026

Revised: 7 June 2026

Accepted: 22 June 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Background

With the rapid advancement of digital technology and the widespread adoption of mobile internet, short-form mobile videos have emerged as a dominant medium for media communication [1–3]. Platforms like TikTok, YouTube, and Douyin have become central to how people consume and share information, offering creative, multimodal content that influences communication, marketing, and public engagement strategies across industries. Short-video platforms have also become important channels for disseminating content related to sustainable development, including environmental protection, cultural heritage preservation, and rural revitalization. By 2023, there were approximately 5 billion social media users globally, with over 1 billion users in China alone, contributing to an overwhelming volume of unstructured and dynamic data. TikTok alone reported over 1.5 billion monthly active

users as of early 2024 [4], with a majority of its audience comprising Generation Z. In the U.S., over 80% of the population uses YouTube [5]. In China, short-video platforms see nearly 80 million new video uploads daily, spanning diverse content types that capture cultural, economic, and societal trends. This scale of content production reflects both the democratization of digital communication and the growing public awareness of sustainability-related issues, from ecological conservation to the promotion of intangible cultural heritage.

The sheer volume of short-video content, however, poses significant challenges for efficiently extracting fine-grained, actionable insights from the vast amounts of multimodal data it generates, particularly for research and governance purposes related to sustainability. These challenges are compounded by the transient and dynamic nature of short videos, which often include rapid scene changes, complex narratives, and diverse symbolic elements such as text overlays, visuals, and soundtracks. Traditional manual methods of analyzing video data are ill-equipped to handle this complexity, being both inefficient and prone to subjectivity [6,7]. The inability to efficiently process and understand such content at scale limits the capacity of researchers, policymakers, and cultural institutions to monitor sustainability narratives, evaluate public communication campaigns, and leverage user-generated content for evidence-based decision-making. While much of the existing research focuses on theoretical analyses or broad summarization techniques, there remains a gap in developing computational methods that enable high-precision extraction and understanding of intricate visual content within short videos [8].

To address this gap, our study proposes a novel method for fine-grained content extraction and understanding in short videos. By leveraging computer vision and multimodal learning models, we aim to dissect and analyze the detailed visual and narrative structure of short videos, focusing on key elements such as scene segmentation, symbolic representations, and temporal continuity. The method is validated on a dataset of culturally and tourism-oriented short videos from China's Douyin platform, closely tied to sustainability goals such as cultural preservation, rural economic development, and the promotion of ecological tourism. The study employs automated video segmentation, keyframe identification, and descriptive text generation to extract detailed semantic information. Furthermore, the study investigates how various interference factors, video content types, and image description models affect the accuracy and expressiveness of the extracted content, offering new insights for automated video analysis and its applications in sustainability-oriented media and communication research. By enabling more efficient and scalable analysis of video data, this work contributes to the broader goal of building intelligent infrastructure that can support sustainable information ecosystems, cultural documentation, and evidence-based public discourse.

2. Related Work

Video processing has become an essential tool in social science research, providing insights into human behavior, public opinion, and media influence [9,10]. Ye et al. [11] developed an emotion prediction model to understand the dynamic structure of video ads by analyzing emotional changes and generating engaging text. Griffin [12] and Peng [8] explored video data collection and automatic visual analysis for social media effects, highlighting the importance of keyword extraction from videos. Fu [13] proposed the WordIllusion algorithm for adversarial text generation, offering insights for keyword extraction from short videos. Bi [14] analyzed motivations behind short-video creation, providing strategic guidance for text generation based on creator intentions. Mourchid [15] introduced a multilayer network model to capture film narratives using scripts, subtitles, and video content. Bu [16] designed a creative process-aware model designed to detect fake news in short videos by analyzing material selection and editing characteristics.

Apostolidis [17] summarized the recent approach to video text summarization. Recent advancements in video summarization methods can be classified into several key categories. These include extractive vs. abstractive summarization; extractive methods select keyframes directly from the video [18–20], while abstractive methods generate new content [21–23]. Supervised, unsupervised, and semi-supervised techniques are also common, with deep learning-based methods such as RNNs, LSTMs, and attention mechanisms becoming popular for handling temporal dependencies and keyframe selection [24,25]. GANs and reinforcement learning have been used to optimize summary quality and diversity [26,27]. Additionally, query-based and context-aware summarization allow for more tailored summaries, while multimodal approaches combine visual, audio, and textual data for comprehensive summaries [28,29]. Lastly, graph-based and hybrid methods offer powerful tools for capturing relationships between frames and improving overall summarization accuracy [30,31].

3. Theoretical Foundation

Our study builds the overall framework based on interrelated strands of communication scholarship that jointly motivate our analytical pipeline: visual narrative theory, multimodal communication (social semiotics), and platform affordance theory. Rather than invoking communication theory in a general sense, we specify how each strand grounds a concrete component of the method: keyframe selection, image-to-text transformation, and the platform-specific interpretation of short-video content, respectively. Firstly, drawing on visual narrative theory, when analyzing the content of short videos, we select keyframes to represent the overall shot content. In short videos, each sub-shot usually constitutes an independent narrative unit that conveys a specific plot or message [32,33]. Successive frames in a video often contain a large amount of redundant information, especially in sections with small content changes [34]. Studies have shown that the human visual system is most sensitive to significant changes in a scene and tends to pay little attention to continuous, similar visual information [35]. Shot switching points are usually accompanied by significant changes in the scene; therefore, frames at these points are more likely to attract the viewer's attention and are regarded as the key content of the shot and keyframes [36,37]. By selecting keyframes, the model is able to effectively capture and retain the core content and major changes of the narrative unit, which in turn ensures the integrity of the overall narrative structure of the short video [38]. In this sense, keyframe extraction is not merely a technical sampling operation but an operationalization of how visual narratives segment meaning into discrete, attention-salient units.

Secondly, the image is converted into a text summary for the extraction of key elements in the short video. This step is grounded in multimodal communication and social semiotics, which hold that meaning is distributed across, and can be re-articulated between, different semiotic resources [39,40]. Although images and text belong to different symbol systems, they are both capable of effectively conveying information and constructing meaning [41]. The conversion between the two symbol systems is mainly manifested in computer language as the recoding of information, i.e., retaining the core meaning of information while adapting it to different communication environments and audience needs, even if it is applicable to a wider range of contexts [42]. Different symbolic modes (e.g., text, image, and sound) have synergistic and transformational relationships in communication [43]. When information is converted from one symbolic mode to another, the core content of the information can still be retained and effectively transmitted despite the change in expression [44]. When these visual elements are extracted and transformed into text, the text becomes another narrative tool, capable of reproducing the narrative structure and core content in the image [37,45]. Framing image-to-text generation as cross-modal transduction, rather than mere captioning, clarifies

why semantic fidelity—not surface lexical overlap—is the appropriate criterion against which the generated descriptions should ultimately be judged.

Thirdly, because our data are drawn from a single short-video platform, we situate the analysis within platform affordance theory [46,47]. Affordances describe the action possibilities that a technological environment makes available to its users, and short-video platforms such as Douyin are characterized by a distinctive set of them: algorithmic recommendation, vertical full-screen viewing, template-based editing, and influencer- and commerce-oriented production. These affordances shape scene composition, editing rhythm, and the visual conventions through which meaning is signaled and therefore directly condition what counts as a salient keyframe and how its content should be described. Recognizing this, we treat our pipeline as a framework tuned to the affordances of platform-native short video, rather than as a platform-neutral method, and we correspondingly limit our empirical claims to the Douyin cultural- and tourism-oriented context studied here.

4. Data Source and Research Methods

Data Source

Our data was sourced from the Douyin platform, which primarily features short-video content related to culture and tourism. To ensure diversity and representativeness, 22 opinion leaders and official accounts with broad influence were selected from Douyin. A total of 8193 videos were collected between January 2022 and December 2024, from which 6554 videos were assigned to the training set (80%), with the remaining 1639 videos held out as the test set (20%). Crucially, this partition was performed at the video level: all frames extracted from a given video were assigned exclusively to either the training or the test set, never split across both. This prevents the data leakage that would arise if near-duplicate frames from the same shot appeared in both partitions. To extract key content, these videos were processed using a video cutting algorithm, yielding approximately 127,021 extracted frames in total (101,509 from the training videos and 25,512 from the test videos). This step was undertaken to capture the keyframes from each video, serving as the foundation for subsequent image annotation.

For the selection of annotators, two staff members with backgrounds in journalism were chosen to ensure the professionalism and accuracy of the annotations. During the annotation process, each annotator independently annotated every image and reviewed the results as needed to ensure consistency. The annotation of image text summaries followed strict criteria: the annotated text had to be completely accurate, covering the main actions, all major objects, and the primary scenes depicted in the image. If no text met these criteria, the annotators were required to select the “all incorrect” option [48]. Before formal annotation, the two coders were trained on a shared codebook that defined each content category and provided positive and negative examples; borderline cases—such as frames containing several competing elements or transitional frames between two shots—were resolved by a documented rule that the dominant element occupying the largest visual area determines the primary category. Disagreements between the two annotators were resolved through discussion, and any case that could not be reconciled was adjudicated by a third senior researcher whose decision was final. Inter-annotator agreement was assessed on a randomly selected subset of 1200 images that both coders annotated independently, and the agreement statistics reported below are computed on this doubly-annotated subset.

Binary coding was employed to categorize the images in our study, ensuring that classification was both concise and systematic, thereby facilitating subsequent data analysis. The code represents five categories from left to right: natural scenery, folk culture, ancient and modern architecture, food, and people. For example, “24” (binary 11,000) indicates that the image belongs to the natural scenery and folk culture categories. Additionally,

to help the model better understand the temporal continuity of video content, the similarity between two neighboring images was also annotated. If two images were visually similar and from the same shot, they were labeled “0”; if they were significantly different or from different shots, they were labeled “1”.

Strict quality control measures were implemented throughout the data annotation process, including two-person annotation, review mechanisms, and sampling checks, to ensure the accuracy and consistency of the data. To evaluate the consistency between the two annotators, Cohen’s kappa was computed separately for each annotation variable. The agreement was $\kappa = 0.83$ for the binary keyframe/duplicate-frame label, $\kappa = 0.81$ for the five-category content-type label, and $\kappa = 0.84$ for the text-summary correctness judgment. All kappa values exceeded the 0.80 threshold conventionally interpreted as substantial-to-almost-perfect agreement [49], demonstrating satisfactory inter-annotator reliability. This ensures the reliability of the annotation process, the trustworthiness of the data, and the scientific rigor of the research findings.

The distribution of image meta-elements and image types is shown in Figure 1, where image types representing less than 1% are not displayed. The proportion of image keyframes is shown in Figure 2.

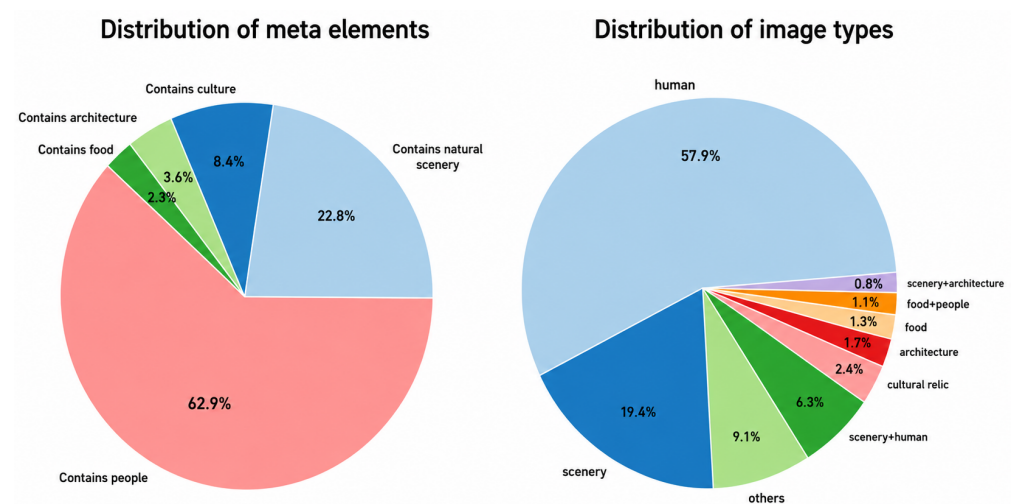


Figure 1. Distribution of image types.

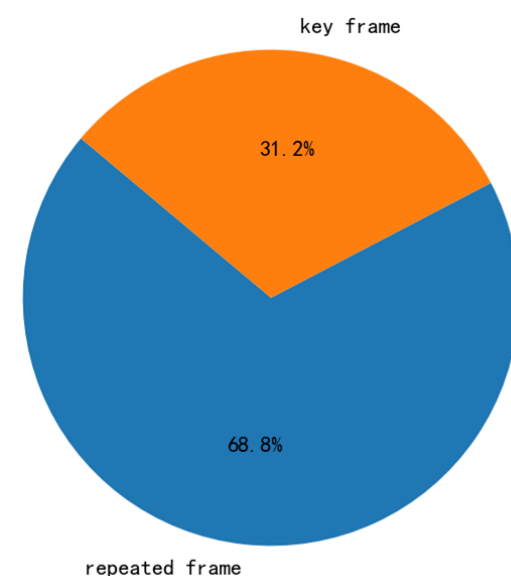


Figure 2. Distribution ratio of keyframes and repeated frames.

5. Methods

5.1. Video Segmentation Algorithm

In the automated processing of videos, to efficiently extract representative keyframes, we propose an algorithmic process based on scene change analysis and similarity filtering. The core idea is to analyze the dynamic characteristics of a video by detecting scene transitions and combining multiple similarity metrics for intelligent sampling of the video content, enabling the extraction of keyframes that retain important visual elements, thereby improving video processing efficiency.

The frequency of camera transitions reflects the rhythm and dynamic changes in the video, which is critical for efficient sampling [50]. We use the PySceneDetect tool to detect scene transitions in the video [51,52]. Specifically, we employ the content-aware detector (ContentDetector), which flags a scene boundary when the weighted change in HSV color space between consecutive frames exceeds a preset threshold. The detection threshold was set to 27.0 (on the 0–255 scale, the tool's default), with a minimum scene length of 15 frames to suppress spurious cuts caused by rapid motion or transient flicker; these values were selected empirically on a held-out subset of 50 videos by comparing automatically detected boundaries against manually identified ones. By analyzing these transition points, the algorithm not only identifies moments where significant content changes occur but also tracks the duration of each shot. This information forms the foundation for subsequent sampling and threshold setting.

After detecting scene transitions, the algorithm determines the duration of the shortest shot in the video. This represents the fastest-changing segment, typically involving scenes with rapid camera movement or quick cuts. To ensure that enough visual information is captured in these fast-paced scenes, we set a sampling threshold k lower than the duration of the shortest shot. In practice, k was set to one-half of the shortest-shot duration (i.e., $k = 0.5 \times d_{\min}$, with a lower bound of $k = 0.5$ s to avoid over-sampling when an extremely short shot was present), which guarantees that at least two frames are sampled from even the briefest shot while limiting redundancy in longer ones. A sensitivity check varying this ratio between 0.3 and 0.7 changed the number of candidate frames by less than 8% and did not materially affect the downstream classification results. Based on threshold k , the algorithm performs intelligent sampling within each shot, ensuring dense frame capture in fast-changing scenes while minimizing the generation of redundant frames in slower, smoother sequences.

5.2. Image Similarity Calculation

After the initial sampling of image frames, the algorithm further filters the sampled frames to remove duplicate frames. This step is achieved by calculating the similarity between neighboring frames. The algorithm uses three commonly used methods for calculating image similarity, namely peak signal-to-noise ratio (PSNR), root mean square error (RMSE), and structural similarity (SSIM). PSNR measures the similarity by comparing the difference in pixel values of an image:

$$\text{PSNR} = 20 \times \log_{10} \left(\frac{\text{MAX}_p}{\sqrt{\text{MSE}}} \right) \quad (1)$$

where MAX_p is the maximum possible value of the pixel values in the image. It is usually 255 for 8-bit images and 4095 for 12-bit images. It can be 1 if the image is of floating-point type and the pixel value is between [0, 1]. MSE represents the mean square error between the original image and the processed image. It measures the difference between the images by calculating the squared mean of the difference in the corresponding pixel values of the two images. The formula is

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (I_1(i) - I_2(i))^2 \quad (2)$$

RMSE reflects the error between images by calculating the square root of the pixel differences:

$$\text{RMSE} = \sqrt{\text{MSE}} \quad (3)$$

SSIM considers the brightness, contrast, and structural information of the images more comprehensively to assess the similarity of the images in multiple dimensions:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (4)$$

where μ_x and μ_y denote the mean values of images x and y , respectively. σ_x^2 and σ_y^2 denote the variance of images x and y , respectively. σ_{xy} denotes the covariance between images x and y . C_1 and C_2 are constants and usually take smaller values to avoid division by zero cases.

Eigenvectors $x = [\text{"PSNR"}, \text{"RMSE"}, \text{"SSIM"}]$ are constructed by calculating the values of the above three metrics between each frame and its neighboring frames. These eigenvectors can adequately describe the differences between frames. Rather than applying fixed cut-off values to each metric independently, the three similarity scores were passed jointly to the SVM classifier described in the next subsection, which learns the decision boundary between near-duplicate and distinct frames from the annotated data. For reference, on the annotated training data, the typical ranges separating the two classes were approximately $\text{PSNR} > 30$ dB, $\text{RMSE} < 10$, and $\text{SSIM} > 0.90$ for frame pairs labeled as near-duplicates; these values are reported only to characterize the data and were not used as hard thresholds.

5.3. SVM-Based Keyframe Classification

To train the SVM model, we frame the keyframe extraction problem as a binary classification task, determining whether a frame is a keyframe or not. Two classes of labels are defined: the positive class (Label = 1) represents keyframes that should be retained—those rich in information and representative of the main content of the current shot; the negative class (Label = 0) represents non-keyframes that should be discarded—those redundant or too similar to neighboring frames. To ensure the model's generalization ability, the training dataset includes a variety of samples with different scenes and degrees of change, and the ratio of positive to negative samples is appropriately balanced.

Data preprocessing is essential before model training. In this study, feature data were normalized by scaling individual feature values to a consistent range (e.g., 0–1) to prevent any feature from disproportionately influencing the model. Additionally, a data cleaning process was employed to remove outliers and noisy data, ensuring the quality and reliability of the training data. The SVM classifier was trained and evaluated using the video-level 8:2 partition described in the Data Source Section (6554 training videos/1639 test videos). Feature vectors derived from all frames within a given video were kept in the same partition so that no frame from a test video was ever seen during training. This ensures that performance evaluation reflects genuine generalization to unseen videos rather than memorization of near-duplicate frames.

After data preprocessing, the SVM model was trained using the training set. The success of an SVM model hinges on selecting the appropriate kernel function and tuning key model parameters (e.g., the penalty parameter C and kernel function parameter γ). The radial basis function (RBF) kernel was chosen for its strong performance in handling

nonlinear problems. Cross-validation was employed to divide the training set into multiple subsets, cycling through them as validation and training sets, to assess the model's performance with different parameter configurations. Grid search was used to explore a range of parameter combinations, identifying those that optimize model performance metrics (e.g., accuracy, precision, and recall). Through these methods, the optimal model parameters were determined, enabling the SVM model to achieve the best classification results on the training set.

The complete SVM configuration was as follows. The feature vector for each frame combined the PSNR, RMSE, and SSIM similarity scores relative to the preceding sampled frame, and all features were standardized to zero mean and unit variance before training. An RBF kernel was used, and the two key hyperparameters were tuned by grid search with 5-fold stratified cross-validation on the training set, optimizing the macro-averaged F1 score. The search grid spanned $C \in \{0.1, 1, 10, 100, 1000\}$ and $\gamma \in \{0.0001, 0.001, 0.01, 0.1, 1\}$, and the optimal configuration was $C = 10$ and $\gamma = 0.01$. Because keyframes (the positive class) were less frequent than duplicate frames, class weights were set inversely proportional to class frequency (the "balanced" scheme), which assigned a higher misclassification cost to the minority keyframe class. To favor high recall for keyframes and thereby reduce the risk of discarding informative frames, the decision threshold on the calibrated posterior probability was lowered from the default 0.5 to 0.35. Probability estimates were obtained through Platt scaling (sigmoid calibration) fitted on the cross-validation folds so that the reported scores reflect calibrated confidence rather than raw decision-function margins. We note that this recall-oriented threshold is a deliberate design choice: it is appropriate for applications such as content monitoring or archival indexing, where missing a salient frame is more costly than retaining a few redundant ones, but it can be tightened toward the default for storage or precision-sensitive applications.

After the SVM model training was completed, the feature vectors of all image frames obtained from sampling were fed into the model to generate classification results. The model classified each frame as either a keyframe (positive class) or a duplicate frame (negative class). Based on the prediction results, frames classified as keyframes were retained, while duplicate frames were discarded.

To further optimize the selection of keyframes, several post-processing strategies were employed. First, a confidence assessment was integrated, allowing the SVM model to output a confidence score for each prediction. Keyframes were then filtered based on confidence levels, ensuring that only frames with high representativeness and informativeness were retained. Additionally, to maintain the temporal continuity of the video, a minimum time interval between keyframes was established. This, along with other strategies, helped prevent keyframes from being either overly concentrated or excessively dispersed, ensuring a more uniform distribution of keyframes along the video timeline.

In our study, misclassifying negative classes as positive results only in increased image summary computation, while misclassifying positive classes as negative leads to the loss of key information. Therefore, special attention is given to reducing false negatives (FNs) and increasing the recall rate. To achieve this, the weight of the positive class was increased in the support vector machine (SVM) model, reducing the risk of misclassifying keyframes as negative. Additionally, the decision threshold was adjusted to make the model more aggressive in classifying frames as positive, which, while reducing FNs, may result in an increase in false positives (FPs).

Through the combined evaluation of the SVMs described above, the algorithm is able to identify image frames with too much similarity and remove them, ultimately retaining the most representative keyframes in each shot, which accurately represent the main visual content in the video. Compared with the original video, keyframes significantly reduce

redundant information, providing a more efficient database for subsequent text generation and content analysis. This intelligent sampling method based on shot switching frequency and similarity filtering is not only able to effectively extract key elements in short videos but may, in principle, be transferable to other types of video processing tasks such as news and education. However, because all data in this study were drawn from a single platform (Douyin) and a single thematic domain (culture and tourism), such cross-domain transferability remains a hypothesis to be tested rather than a demonstrated property, and we therefore refrain from claiming that the method is universally applicable. In addition, our algorithm achieves efficient segmentation and keyframe extraction of video content, which significantly improves processing efficiency while ensuring the integrity of video content, and innovatively achieves intelligent processing and analysis of video content.

5.4. Keyframe Image Description Algorithm

We employed a keyframe image description generation method based on the mPLUG model, a multimodal learning framework capable of processing and generating both image and text information simultaneously [53]. Specifically, we adopted the Chinese-language mPLUG variant (mPLUG-base) released on the ModelScope platform, whose visual encoder is a ViT-B/16 backbone and whose text decoder was pre-trained on large-scale Chinese image–text corpora, making it well-suited to the Chinese cultural–tourism content in our dataset. Rather than using the model in a zero-shot setting, we fine-tuned it on the annotated keyframe–caption pairs from the training videos for 10 epochs using the AdamW optimizer (learning rate 2×10^{-5} , batch size 16, maximum sequence length 64 tokens), selecting the checkpoint with the best validation CIDEr score. Each training image was paired with the single human-authored reference description produced during annotation; the held-out evaluation images were each assigned three independent reference descriptions to support more reliable automatic scoring. The input to the model was the raw keyframe image together with a fixed Chinese task prompt instructing the model to describe the main scene, objects, and actions. To mitigate hallucination, generation used beam search (beam size 5) with a repetition penalty of 1.2 and a no-repeat- n -gram constraint, and any generated description whose objects could not be grounded in the image during a manual spot-check was flagged and excluded from the reported results. The model effectively extracts rich semantic information from the keyframes of the video and accurately converts it into textual descriptions through multilevel cross-modal fusion. As illustrated in Figure 3, the mPLUG model comprises two independent unimodal encoders, a network of cross-modal skip connections, and a decoder for text generation.

The unimodal encoder component of the mPLUG model processes the image information and the corresponding text information from the video keyframes separately. The image encoder utilizes a Vision Transformer architecture, which chunks the image into patches and encodes these patches into a serialized embedded representation. Unlike traditional visual feature extraction methods, the Vision Transformer can process image data more efficiently while preserving richer visual features. Additionally, each image embedding sequence includes a special [CLS] token, which summarizes global information about the entire image, providing a global context for subsequent cross-modal processing. Simultaneously, the text encoder processes the input text data to generate a text embedding sequence, analogous to the image embedding process, and similarly uses the [CLS] token to summarize the textual information.

After the unimodal encoding is complete, the image and text embedding representations are fed into the cross-modal skip connection network of the mPLUG model. This network is built around multiple cross-modal fusion blocks, each containing several asymmetric co-attention layers. These blocks facilitate deep fusion of multimodal information

through the skip connection mechanism. The core concept of cross-modal fusion is that the model can effectively combine features from different modalities using an asymmetric co-attention mechanism, allowing it to preserve the uniqueness of each modality while enhancing the overall representation of the information. By recursively applying these cross-modal fusion blocks, the model progressively improves its ability to capture complex relationships between images and text, resulting in more accurate and semantically rich cross-modal representations for subsequent text generation.

The fused cross-modal representations are then fed into the mPLUG model's decoder, which is built on the Transformer architecture and is primarily designed for sequence-to-sequence text generation tasks. The decoder utilizes a combination of causal self-attention and cross-attention mechanisms, allowing the model to fully leverage the previously fused cross-modal information when generating textual descriptions. This design ensures coherence and contextual consistency throughout the text generation process. As a result, mPLUG is capable of not only generating accurate text summaries that reflect the content of the keyframes but also preserving the rich semantic information contained in the images, leading to highly descriptive and precise generated text.

Ultimately, the initial text generated by the mPLUG model undergoes a series of post-processing steps to ensure coherence and readability. First, the text is filtered to remove any redundant information or repetitive descriptions, ensuring that it remains concise and information-dense. Next, the text is processed using Jieba segmentation, followed by de-duplication of stopwords, converting the text into lemmas for further analysis or use.

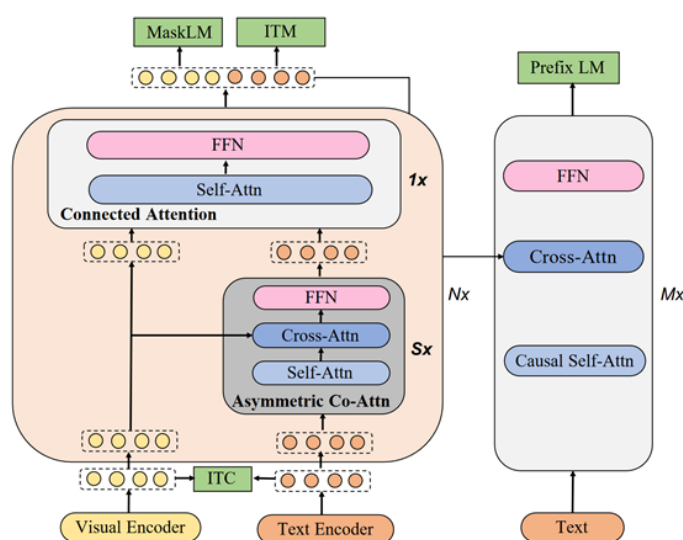


Figure 3. mPLUG process.

6. Results

6.1. Results of Keyframe Extraction

Figure 4 shows the relationship between the categories and the three performance metrics (SSIM, RMSE, and PSNR), where the red dots represent category 0 and the blue dots represent category 1. The boundary is relatively clear, indicating the possibility of binary classification.

Figure 5 shows the learning curve of the support vector machine (SVM) model with different sizes of training samples. The horizontal axis represents the number of samples in the training set, while the vertical axis indicates the classification accuracy of the model. The red curve represents the accuracy on the training set, and the green curve represents the accuracy on the cross-validation set. From the learning curve, it can be observed that the SVM model's training accuracy and cross-validation accuracy begin to stabilize

when the number of training samples reaches approximately 4500. When the number of samples exceeds 6000, the training and validation accuracy curves show a consistent trend, indicating that the model's performance is stable and its generalization ability is good.

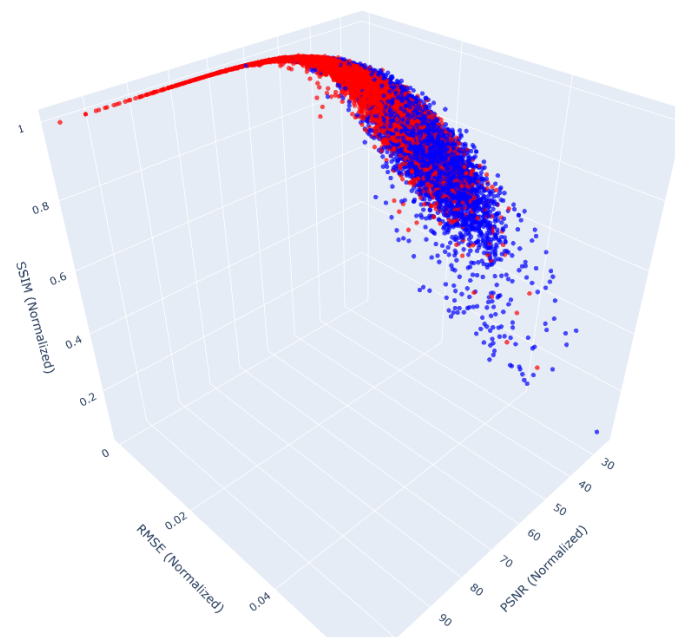


Figure 4. Relationship between SSIM, RMSE, PSNR, and categories.

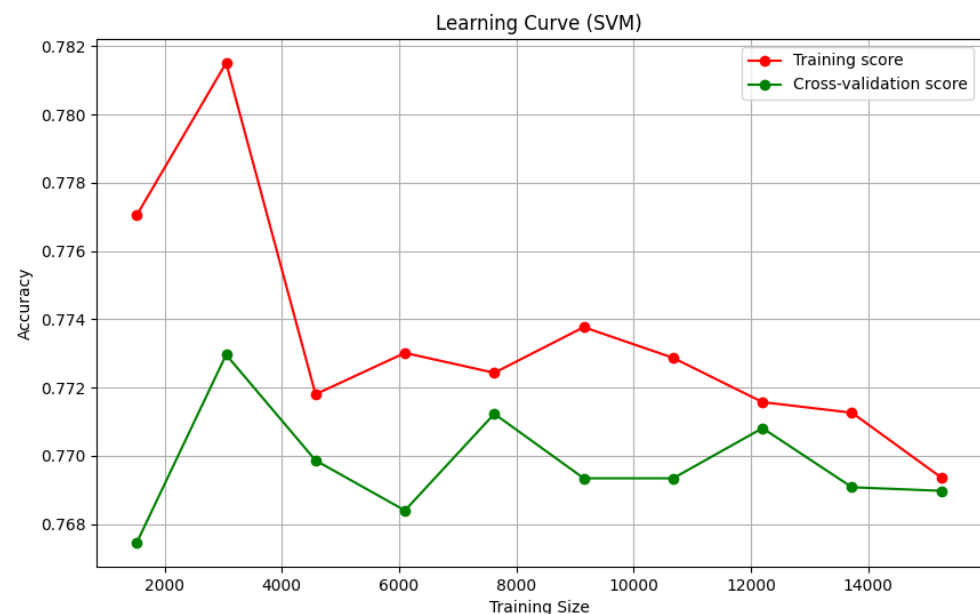


Figure 5. Learning curve of the support vector machine (SVM) model with different training sample sizes.

To evaluate the effectiveness of the SVM model in the task of keyframe extraction, performance testing was conducted using the test set. The classification performance of the model was comprehensively evaluated using metrics such as accuracy, precision, recall, and F1-score (as shown in Table 1), along with the confusion matrix (as shown in Figure 6).

In the SVM model's test results, the model demonstrated good overall classification performance, with an overall accuracy of 82%. In terms of classification precision, the model achieved a precision of 95% for class 0 and 66% for class 1. It is important to note that the lower precision for class 1 was intentional to reduce false negatives (FNs) and in-

crease the recall rate for the positive class (class 1), thereby minimizing the risk of missing keyframes and losing critical information. This strategy aims to maximize the identification of keyframes, even though it may introduce more false positives (FPs), where negative samples are incorrectly classified as positive.

The recall results highlight the model's sensitivity in identifying the positive class (class 1), prioritizing keyframe identification, even at the expense of some accuracy. This strategy is crucial in video processing tasks, particularly in keyframe extraction, where avoiding the omission of important information is essential. The F1-score further reflects the model's balance between precision and recall, indicating that despite the lower precision in identifying positive samples, the overall performance remains balanced, aligning with the study's emphasis on minimizing false negatives.

Table 1. Classification performance metrics of the SVM model.

Category	Precision	Recall	F1-Score	Count
0	0.95	0.77	0.85	1570
1	0.66	0.91	0.77	774
Accuracy			0.82	2344
Macro Average	0.80	0.84	0.81	2344
Weighted Average	0.85	0.82	0.82	2344

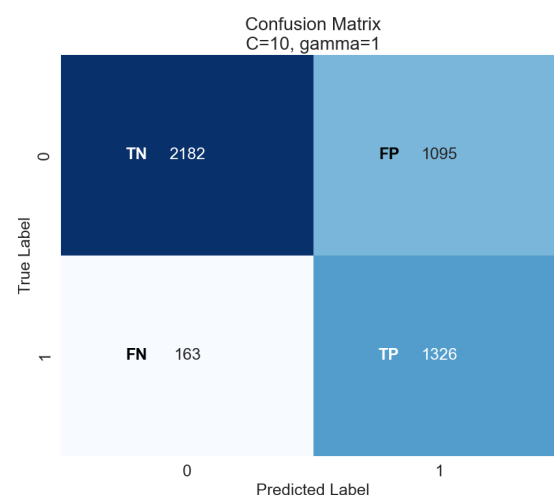


Figure 6. Confusion matrix of SVM on the test set.

6.2. Results of Image Description

6.2.1. Results of the Image Description Across Different Categories

As shown in Figure 7, the image description algorithm produces good text summaries.

We evaluated the quality of text generation for image descriptions using metrics such as BLEU [54], BertScore [55], SPICE [56], CIDEr [57], and CHR-F [58] (shown in Table 2). Each metric assesses text similarity from different levels, providing a comprehensive analysis of the generated text. BertScore showed high semantic consistency for descriptions involving people, while the score was lower for natural scenery due to the diversity of descriptions. SPICE and CIDEr also performed well in capturing semantic content for people and architecture, but lower scores were noted for natural scenery and folk culture due to their complexity. BLEU scores were low, attributed to the flexible grammar and word order in Chinese, making n-gram matching more challenging. CHR-F scores, particularly suitable for Chinese, also indicated room for improvement in character-level matching. Overall, the model excelled in categories with more focused semantic content, like people

and architecture, but struggled with categories requiring more descriptive diversity, such as natural scenery and food.



Figure 7. Results of the image description algorithm across different categories.

Table 2. Results of various metrics across different categories.

Metric	Natural Scenery	Folk Culture	Architecture	Food	People	Overall
BertScore-P	0.6427	0.6885	0.7902	0.7665	0.8514	0.7901
BertScore-R	0.7004	0.7032	0.8093	0.8169	0.8611	0.8288
BertScore-F1	0.6779	0.6995	0.7844	0.7902	0.8536	0.8149
SPICE	0.5972	0.6038	0.7321	0.6552	0.7946	0.7498
BLEU	0.4516	0.5254	0.6312	0.5014	0.7210	0.6662
CIDEr	4.8384	5.0720	6.1496	5.3998	7.0055	6.5043
CHR-F	0.6020	0.6223	0.7255	0.6459	0.8007	0.7519

6.2.2. Results of the Image Description Under Different Interference Conditions

We examined the impact of different interference factors on the performance of image classification. A total of six types of interference were defined. As shown in Figure 8, (a) represents portrait mode without interference, (b) represents portrait mode with solid color border interference, (c) represents portrait mode with translucent border interference, (d) represents portrait mode with image border interference, (e) represents landscape mode without interference, and (f) represents floating text interference.

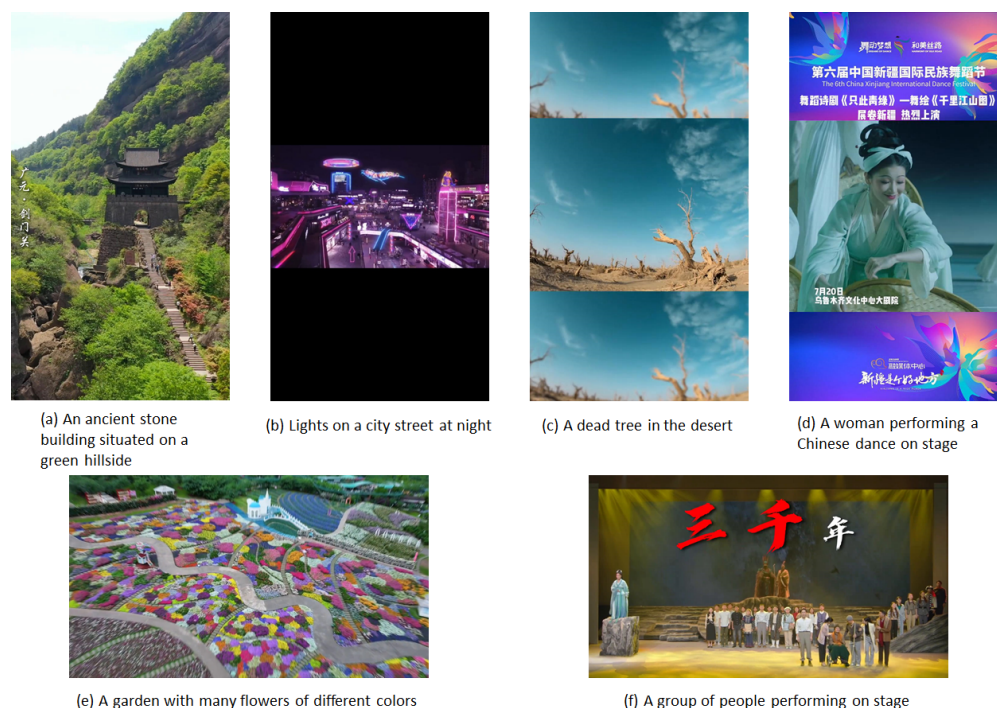


Figure 8. Partial results of the image description algorithm without interference.

The experimental results are shown in Table 3, indicating significant differences in the impact of various types of interference on image classification performance. In the absence of any interference, the model performed well in both portrait and landscape images, with landscape mode without interference performing slightly better than portrait mode without interference. This is because landscape images provide a wider field of view, allowing the model to capture more contextual information.

Table 3. Results of various metrics for the image description algorithm under different interference conditions.

Metric	Landscape No Interference	Portrait No Interference	Portrait Translucent Border	Portrait Solid Color Border	Portrait Image Border	Floating Text Interference
BertScore-P	0.8039	0.8002	0.6029	0.6856	0.6231	0.6984
BertScore-R	0.8226	0.8212	0.6395	0.7284	0.6180	0.6526
BertScore-F1	0.8157	0.8109	0.5853	0.6945	0.5916	0.6640
SPICE	0.7503	0.7496	0.6300	0.6993	0.6792	0.6582
BLEU	0.6601	0.6748	0.5475	0.6238	0.5319	0.5991
CIDEr	6.6859	6.5822	4.7635	5.7735	4.8914	5.4382
CHR-F	0.7554	0.7525	0.6584	0.7358	0.6115	0.7133

When interference is present, classification performance decreases, particularly when borders are added to the images. The metrics significantly dropped under the condition of portrait mode with translucent border interference, as the translucent border increased the model's confusion between the main subject and the background, making it difficult for the model to accurately distinguish key features in the image. Similarly, performance decreased under the conditions of portrait mode with solid color borders and image borders, indicating that although the borders themselves do not directly obscure image content, they still reduce the amount of useful image information, leading to a decline in classification performance.

The interference caused by floating text has a more complex impact on the model's performance. Experimental results show that while the impact of floating text on some metrics is slightly lower compared to border interference, it has a more significant effect on metrics such as CIDEr and BLEU. This effect likely stems from the model misinterpreting large floating text as a "sign" in the image or performing simple text recognition, leading to substantial semantic misunderstandings. For example, the model may mistake the floating text in the image as part of the main content, and this effect becomes more pronounced when the text covers a large area. As a result, the presence of floating text tends to disrupt the model's correct interpretation of the image's semantics, particularly in tasks related to text generation and word frequency matching. This misrecognition can lead to significant deviations, which in turn affect the performance of semantic and word frequency-related metrics, such as SPICE and CIDEr.

6.2.3. Results of Different Image Description Algorithms

Table 4 presents the performance of different image description algorithms across various evaluation metrics, where those marked with "*" represent the results of Chinese image description algorithms translated into English. To ensure a controlled comparison, the four image description models (mPLUG (Chinese), BLIP, mPLUG (English), and vit-gpt-diffusion-zh) were all evaluated on the same keyframes produced by our pipeline (scene segmentation, similarity-based filtering, and SVM selection), with the same single-frame image input; only the final description generation module differs between them. All models were used in a zero-shot setting without task-specific fine-tuning. This isolates the effect of the description model itself, while we note that the English-language models (BLIP, mPLUG (English), and vit-gpt2) are evaluated on machine-translated output, which is a limitation of this part of the comparison. Fine-grained attribute coverage calculates the percentage of elements identified by the algorithm compared to the total number of elements present in the video. For example, if a video contains 20 distinguishable elements (e.g., objects, actions, and scenes) and the algorithm identifies 18, the fine-grained attribute coverage would be 90%. This metric highlights the algorithm's ability to comprehensively extract details from videos. The results indicate that the mPLUG (Chinese) algorithm achieves the strongest scores across these metrics among the models compared under this shared-pipeline setting. The BLIP [59] and vit-gpt2 [60] image description algorithms show a decline in semantic-related metrics due to the loss during the translation process into Chinese. The vit-gpt-diffusion-zh, which is also a Chinese algorithm, lags behind in all metrics, suggesting that its poor performance in the image description task may be related to the specificity of its model architecture and pretraining data.

When compared to algorithms like S2VT [61] and LTX-Video, which focus on video captioning through sequence-to-sequence models, the proposed algorithm demonstrates stronger performance in fine-grained content understanding. Unlike the image description models above, S2VT and LTX-Video are end-to-end video-to-text models that take the complete video clip as input rather than the keyframes extracted by our pipeline; their input type therefore differs, and they are not directly comparable to the shared-pipeline models. For this reason, they are reported only on fine-grained attribute coverage, the single metric that is meaningful across both settings, and the image-specific metrics are marked "-" as not applicable. These video-to-text models primarily generate captions by modeling temporal sequences in videos but lack the capability to comprehensively extract and analyze detailed visual elements within individual frames. In contrast, the proposed method excels in identifying intricate details such as objects, actions, and scene composition, as evidenced by its high fine-grained attribute coverage.

Table 4. Comparison of metrics for different image description algorithms. Models marked “*” produce Chinese output evaluated after translation into English. The first four models share our keyframe extraction pipeline and identical single-frame input, differing only in the description module; all models are zero-shot. S2VT and LTX-Video are end-to-end video-to-text models with full-video input and are therefore reported only on fine-grained attribute coverage; “-” denotes metrics that are not applicable to this setting.

Metric	mPLUG * (Chinese)	BLIP	mPLUG (English)	vit-gpt2	vit-gpt-diffusion-zh *	S2VT	LTX-Video
BertScore-P	0.7901	0.7832	0.6811	0.6742	0.4813	-	-
BertScore-R	0.8288	0.8096	0.6975	0.6994	0.5072	-	-
BertScore-F1	0.8149	0.8070	0.6703	0.6901	0.4985	-	-
SPICE	0.7498	0.6948	0.6210	0.6527	0.5402	-	-
BLEU	0.6662	0.6693	0.5876	0.5825	0.5548	-	-
CIDEr	6.5043	6.3715	5.5715	5.3918	0.9001	-	-
CHR-F	0.7519	0.7045	0.7160	0.6874	0.5291	-	-
Fine-Grained Attribute Coverage	0.9322	0.8517	0.8921	0.8274	0.7759	0.8126	0.8518

7. Conclusions

We propose a method for fine-grained content extraction and text description generation tailored for short videos. Using a short-video dataset of literature and tourism collected and annotated from the Douyin platform, we conducted systematic experimental validation. The results demonstrate the method’s effectiveness in accurately extracting key elements and generating semantically rich descriptions across the diverse cultural- and tourism-related scenarios represented in our Douyin dataset. Furthermore, this study provides a comprehensive analysis of the method’s performance under different conditions, offering valuable insights into its strengths and limitations and paving the way for more precise automated analysis and application of short-video content.

Looking ahead, future work will prioritize enhancing the algorithm’s precision and robustness. This includes addressing visual interference more effectively, developing advanced preprocessing techniques to automatically identify and mitigate irrelevant visual noise, and improving the descriptive richness and semantic accuracy of text generation, particularly in complex categories like landscapes. The study will also examine the role of visual attributes (e.g., brightness and color) in influencing text generation and explore adaptive image enhancement strategies to optimize results. Efforts will focus on strengthening the model’s multimodal learning capabilities by integrating diverse data streams (e.g., speech, text, and visuals) to achieve more holistic analysis. Additionally, we aim to expand the model’s applicability to various video types, including other short videos and long-form documentaries, in order to assess whether consistent and scalable performance can be achieved beyond the present setting and, where possible, to work toward broadening its practical utility and adaptability across diverse media contexts.

Author Contributions: Conceptualization, Y.W. and S.Z.; methodology, Y.W. and S.Z.; software, Y.W.; validation, K.Z., Y.X. and H.W.; formal analysis, Y.W. and S.Z.; investigation, S.Z., K.Z., Y.X. and H.W.; resources, S.Z.; data curation, K.Z., Y.X. and H.W.; writing—original draft preparation, Y.W. and S.Z.; writing—review and editing, Y.W., S.Z. and M.L.; visualization, Y.W.; supervision, M.L.; project administration, M.L.; funding acquisition, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Philosophy and Social Science Research Project of Jilin University, “Research on Enhancing the Effectiveness and Innovating the Pathways of Chinese Civilization Communication in the Digital-Intelligence Era” (Grant No. 2025SZQHYB03).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are not publicly available due to privacy, ethical, copyright, and platform policy restrictions concerning social media and user-generated content. Access to the data is restricted, and any request for access will be considered by the corresponding author on a case-by-case basis, subject to applicable ethical and legal requirements.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Shutsko, A. User-generated short video content in social media. A case study of TikTok. In *Proceedings of the Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, 19–24 July 2020*; Proceedings, Part II 22; Springer: Berlin/Heidelberg, Germany, 2020; pp. 108–125.
- Kaye, D.B.V.; Chen, X.; Zeng, J. The co-evolution of two Chinese mobile short video apps: Parallel platformization of Douyin and TikTok. *Mob. Media Commun.* **2021**, *9*, 229–253.
- Yang, S.; Zhao, Y.; Ma, Y. Analysis of the reasons and development of short video application-Taking Tik Tok as an example. In *Proceedings of the 2019 9th International Conference on Information and Social Science (ICISS 2019)*, Manila, Philippines, 12–14 July 2019; pp. 12–14.
- Statista. Social Media-Statistics & Facts. 2024. Available online: <https://www.statista.com/topics/1164/social-networks/#topicOverview> (accessed on 16 July 2024).
- Auxier, B.; Anderson, M. Social media use in 2021. *Pew Res. Cent.* **2021**, *1*, 1–4.
- Chen, T.S.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.w.; Jeon, B.E.; Fang, Y.; Lee, H.Y.; Ren, J.; Yang, M.H.; et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 24–27 June 2024; pp. 13320–13331.
- Song, H.; Tolochko, P.; Eberl, J.M.; Eisele, O.; Greussing, E.; Heidenreich, T.; Lind, F.; Galyga, S.; Boomgaarden, H.G. In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Commun.* **2020**, *37*, 550–572. [\[CrossRef\]](#)
- Peng, Y.; Lock, I.; Ali Salah, A. Automated visual analysis for the study of social media effects: Opportunities, approaches, and challenges. *Commun. Methods Meas.* **2024**, *18*, 163–185.
- Nyhuis, D.; Ringwald, T.; Rittmann, O.; Gschwend, T.; Stiefelwagen, R. Automated video analysis for social science research 1. In *Handbook of Computational Social Science, Volume 2*; Routledge: London, UK, 2021; pp. 386–398.
- Nassauer, A.; Legewie, N.M. *Video Data Analysis: How to Use 21st Century Video in the Social Sciences*; Sage: Thousand Oaks, CA, USA, 2022.
- Ye, K.; Buettner, K.; Kovashka, A. Story understanding in video advertisements. *arXiv* **2018**, arXiv:1807.11122.
- Griffin, T. A discussion of video as a data collection tool. *Curr. Issues Tour.* **2019**, *22*, 2183–2196.
- Fu, H.; Wang, C.; Sun, J.; Zhao, Y.; Lin, H.; Sun, J.; Zhang, B. WordIllusion: An adversarial text generation algorithm based on human cognitive system. *Cogn. Syst. Res.* **2024**, *83*, 101179. [\[CrossRef\]](#)
- Bi, X.; Tang, C. Research on the motives affecting the behavior of short video’s creators. *IEEE Access* **2020**, *8*, 188415–188428. [\[CrossRef\]](#)
- Mourchid, Y.; Renoust, B.; Roupin, O.; Văn, L.; Cherifi, H.; Hassouni, M.E. Movienet: A movie multilayer network model using visual and textual semantic cues. *Appl. Netw. Sci.* **2019**, *4*, 121. [\[CrossRef\]](#)
- Bu, Y.; Sheng, Q.; Cao, J.; Qi, P.; Wang, D.; Li, J. FakingRecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process. *arXiv* **2024**, arXiv:2407.16670.
- Apostolidis, E.; Adamantidou, E.; Metsai, A.I.; Mezaris, V.; Patras, I. Video summarization using deep neural networks: A survey. *Proc. IEEE* **2021**, *109*, 1838–1863. [\[CrossRef\]](#)
- Pandian, A.A.; Maheswari, S. A keyframe selection for summarization of informative activities using clustering in surveillance videos. *Multimed. Tools Appl.* **2024**, *83*, 7021–7034.
- Yasmin, G.; Chowdhury, S.; Nayak, J.; Das, P.; Das, A.K. Key moment extraction for designing an agglomerative clustering algorithm-based video summarization framework. *Neural Comput. Appl.* **2023**, *35*, 4881–4902.
- Parikh, V.; Mehta, J.; Shah, S.; Sharma, P. Comparative analysis of keyframe extraction techniques for video summarization. *Recent Adv. Comput. Sci. Commun. (Former. Recent Patents Comput. Sci.)* **2021**, *14*, 2761–2771. [\[CrossRef\]](#)
- Hsu, T.C.; Liao, Y.S.; Huang, C.R. Video summarization with spatiotemporal vision transformer. *IEEE Trans. Image Process.* **2023**, *32*, 3013–3026. [\[CrossRef\]](#) [\[PubMed\]](#)

22. Bao, H.; Dong, L.; Wang, W.; Yang, N.; Piao, S.; Wei, F. Fine-tuning pretrained transformer encoders for sequence-to-sequence learning. *Int. J. Mach. Learn. Cybern.* **2024**, *15*, 1711–1728.
23. Feng, S.; Xie, Y.; Wei, Y.; Yan, J.; Wang, Q. Transformer-based video summarization with spatial-temporal representation. In *Proceedings of the 2022 8th International Conference on Big Data and Information Analytics (BigDIA)*; IEEE: Piscataway, NJ, USA, 2022; pp. 428–433.
24. Liu, Y.T.; Li, Y.J.; Wang, Y.C.F. Transforming multi-concept attention into video summarization. In *Proceedings of the Asian Conference on Computer Vision*, Kyoto, Japan, 30 November–4 December 2020.
25. Li, P.; Ye, Q.; Zhang, L.; Yuan, L.; Xu, X.; Shao, L. Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognit.* **2021**, *111*, 107677. [\[CrossRef\]](#)
26. Yuan, Y.; Zhang, J. Unsupervised video summarization via deep reinforcement learning with shot-level semantics. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 445–456. [\[CrossRef\]](#)
27. Apostolidis, E.; Adamantidou, E.; Metsai, A.I.; Mezaris, V.; Patras, I. AC-SUM-GAN: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 3278–3292.
28. Li, H.; Zhu, J.; Ma, C.; Zhang, J.; Zong, C. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 7–11 September 2017; pp. 1092–1102.
29. Huang, J.H.; Murn, L.; Mrak, M.; Worring, M. Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, Taipei, Taiwan, 16–19 November 2021; pp. 580–589.
30. Zhu, W.; Han, Y.; Lu, J.; Zhou, J. Relational reasoning over spatial-temporal graphs for video summarization. *IEEE Trans. Image Process.* **2022**, *31*, 3017–3031. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Chai, C.; Lu, G.; Wang, R.; Lyu, C.; Lyu, L.; Zhang, P.; Liu, H. Graph-based structural difference analysis for video summarization. *Inf. Sci.* **2021**, *577*, 483–509. [\[CrossRef\]](#)
32. Sabha, A.; Selwal, A. Data-driven enabled approaches for criteria-based video summarization: A comprehensive survey, taxonomy, and future directions. *Multimed. Tools Appl.* **2023**, *82*, 32635–32709. [\[CrossRef\]](#)
33. Cao, R.; Dey, S.; Cunningham, A.; Walsh, J.; Smith, R.T.; Zucco, J.E.; Thomas, B.H. Examining the use of narrative constructs in data videos. *Vis. Inform.* **2020**, *4*, 8–22. [\[CrossRef\]](#)
34. Tu, Z.; Yu, X.; Wang, Y.; Birkbeck, N.; Adsumilli, B.; Bovik, A.C. RAPIQUE: Rapid and accurate video quality prediction of user generated content. *IEEE Open J. Signal Process.* **2021**, *2*, 425–440. [\[CrossRef\]](#)
35. Lee, S.M.; Shin, J.; Lee, I. Significance of visual scene-based learning in the hippocampal systems across mammalian species. *Hippocampus* **2023**, *33*, 505–521. [\[PubMed\]](#)
36. Smith, T.J. The attentional theory of cinematic continuity. *Projections* **2012**, *6*, 1–27. [\[CrossRef\]](#)
37. Loschky, L.C.; Larson, A.M.; Smith, T.J.; Magliano, J.P. The scene perception & event comprehension theory (SPECT) applied to visual narratives. *Top. Cogn. Sci.* **2020**, *12*, 311–351. [\[PubMed\]](#)
38. Xu, X.; Yang, L.; Yip, D.; Fan, M.; Wei, Z.; Qu, H. From ‘wow’ to ‘why’: Guidelines for creating the opening of a data video with cinematic styles. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, New Orleans, LA, USA, 30 April–5 May 2022; pp. 1–20.
39. Kress, G.; Van Leeuwen, T. *Reading Images: The Grammar of Visual Design*; Routledge: London, UK, 2020.
40. Kress, G. *Multimodality: A Social Semiotic Approach to Contemporary Communication*; Routledge: London, UK, 2009.
41. Barthes, R. *Elements of Semiology*; Jonathan Cape: London, UK, 1968; Volume 4.
42. Bateman, J.A. What are digital media? *Discourse Context Media* **2021**, *41*, 100502. [\[CrossRef\]](#)
43. Pederson, R.; Myers, J. Critical literacy and semiotic juxtaposition: The possibilities of multimodal pedagogy in Korea. *J. Engl. Teach. Through Movies Media* **2021**, *22*, 73–92. [\[CrossRef\]](#)
44. Chen, W.; Wang, W.; Liu, L.; Lew, M.S. New ideas and trends in deep multimodal content understanding: A review. *Neurocomputing* **2021**, *426*, 195–215. [\[CrossRef\]](#)
45. Nielsen, W.; Turney, A.; Georgiou, H.; Jones, P. Meaning making with multiple representations: A case study of a preservice teacher creating a digital explanation. *Res. Sci. Educ.* **2022**, *52*, 871–890. [\[CrossRef\]](#)
46. Gibson, J.J. *The Ecological Approach to Visual Perception: Classic Edition*; Psychology Press: New York, NY, USA, 2014.
47. Bucher, T.; Helmond, A. The affordances of social media platforms. *SAGE Handb. Soc. Media* **2018**, *1*, 233–253. [\[CrossRef\]](#)
48. Wevers, M.; Vriend, N.; De Bruin, A. What to do with 2,000,000 historical press photos? The challenges and opportunities of applying a scene detection algorithm to a digitised press photo collection. *TMG J. Media Hist.* **2022**, *25*, 1–24. [\[CrossRef\]](#)
49. Landis, J.R.; Koch, G.G. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* **1977**, *33*, 363–374. [\[CrossRef\]](#) [\[PubMed\]](#)
50. e Souza, M.R.; de Almeida Maia, H.; Vieira, M.B.; Pedrini, H. Survey on visual rhythms: A spatio-temporal representation for video sequences. *Neurocomputing* **2020**, *402*, 409–422. [\[CrossRef\]](#)

51. PySceneDetect. Available online: <https://github.com/Breakthrough/PySceneDetect/> (accessed on 10 June 2024).
52. Hussain, Z.; Zhang, M.; Zhang, X.; Ye, K.; Thomas, C.; Agha, Z.; Ong, N.; Kovashka, A. Automatic understanding of image and video advertisements. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1705–1715.
53. Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv* **2022**, arXiv:2205.12005.
54. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
55. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.
56. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Spice: Semantic propositional image caption evaluation. In *Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Proceedings, Part V 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 382–398.
57. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575.
58. Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, 17–18 September 2015; pp. 392–395.
59. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2022; pp. 12888–12900.
60. Kumar, A. The Illustrated Image Captioning Using Transformers. *Ankur | NLP Enthusiast*, 2022. Available online: <https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers/> (accessed on 21 June 2026).
61. Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; Saenko, K. Sequence to sequence-video to text. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4534–4542.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.