

VIDEO EVALUATION METHOD AND ELECTRONIC DEVICE

Technical Field

[0001] The present application belongs to a field of artificial intelligence technology, relates to deep learning technology, and in particular to a video evaluation method and electronic device.

Background Art

[0002] With a development of artificial intelligence generated content (AIGC) technology, AIGC videos are becoming increasingly widely used in e-commerce. However, quality of AIGC videos varies widely. An existing method for evaluating the AIGC video lacks an understanding of hierarchical dynamic features of a video and has limited generalization capabilities, making it difficult to obtain efficient and accurate evaluation results for AIGC videos in e-commerce scenarios.

Summary

[0003] The embodiments of the present application provide a video evaluation method and an electronic device, which can solve the existing problem of difficulty in obtaining efficient and accurate evaluation results of AIGC videos in e-commerce scenarios.

[0004] A first aspect of an embodiment of the present application provides a video evaluation method, comprising: sampling a video based on a plurality of sampling rates and obtaining a plurality of sub-videos of the video; obtaining a first feature vector corresponding to each of the plurality of sub-videos by extracting features from each of the plurality of sub-videos using a video evaluation model; obtaining a merged feature vector by merging first feature vectors corresponding to all of the plurality of sub-videos; obtaining a second feature vector by extracting features the merged feature vector using the video evaluation model; and obtaining an evaluation result of the video by evaluating the video using the video evaluation model according to preset prompt information and the second feature vector.

[0005] According to an embodiment of the present application, sampling the video based on the plurality of sampling rates to obtain the plurality of sub-videos of the video comprises: determining the plurality of sampling rates according to a frame rate of the video; obtaining a

video frame sequence corresponding to each of the plurality of sampling rates by extracting video frames from the video according to each of the plurality of sampling rates; and obtaining a sub-video corresponding to each of the plurality of sampling rates according to the video frame sequence corresponding to each of the plurality of sampling rates.

5 **[0006]** According to an embodiment of the present application, the video evaluation model comprises a visual encoder, and wherein obtaining the first feature vector corresponding to each of the plurality of sub-videos by extracting features from each of the plurality of sub-videos using the video evaluation model comprises: obtaining a feature vector corresponding to each video frame of each of the plurality of sub-videos by extracting image features from each video
10 frame using the visual encoder; and merging the feature vectors corresponding to all video frames of each of the plurality of sub-videos, and obtaining the first feature vector corresponding to each of the plurality of sub-videos.

[0007] According to an embodiment of the present application, obtaining the merged feature vector by merging the first feature vectors corresponding to all of the plurality of sub-videos comprises:
15 obtaining the merged feature vector by merging the first feature vectors corresponding to all of the plurality of sub-videos according to a channel splicing technology.

[0008] According to an embodiment of the present application, the video evaluation model comprises a multi-layer perceptron, and obtaining the second feature vector by extracting features from the merged feature vector using the video evaluation model comprises: performing a nonlinear
20 transformation on the merged feature vector using the multi-layer perceptron, and obtaining a nonlinear transformation result; compressing a dimensionality of the nonlinear transformation result to a preset dimensionality, and obtaining the second feature vector.

[0009] According to an embodiment of the present application, the preset prompt information comprises first prompt information corresponding to the video, the video is an artificial intelligence-
25 generated video generated based on the first prompt information.

[0010] According to an embodiment of the present application, the preset prompt information comprises second prompt information corresponding to the video, and the second prompt information is used to instruct the video evaluation model to perform at least one preset evaluation operation on the video, and the at least one preset evaluation operation comprises one or more of the following:

evaluating a degree of semantic alignment between the video and the first prompt information;
evaluating a temporal quality of the video, where the temporal quality comprises a combination of
one or more of a jitter degree, a flicker degree, and a coherence degree between adjacent video frames
of the video; evaluating an image quality of the video frames of the video, wherein the image quality
5 comprises a clarity level and/or a noise level; evaluating an aesthetic quality of the video frames of
the video, where the aesthetic quality comprises a content stability level and/or a color harmony level.

[0011] According to an embodiment of the present application, the video evaluation model
comprises a large language model, and obtaining the evaluation result of the video by evaluating the
video using the video evaluation model according to preset prompt information and the second feature
10 vector comprises: based on the second prompt information, performing any one of the evaluation
operations on the video using the large language model, and obtaining a corresponding evaluation
result.

[0012] According to an embodiment of the present application, the video evaluation method further
comprises: training the video evaluation model using a pre-training fine-tuning method and a preset
15 loss function until the preset loss function meets a preset condition.

[0013] A second aspect of the present application provides an electronic device, comprising:
storage device, and a processor, wherein the processor executes computer-readable instructions stored
in the storage device to implement the video evaluation method.

[0014] The video evaluation method provided in the embodiments of the present application can
20 generate a plurality of sub-videos based on a plurality of sampling rates, thereby facilitating the
capture of dynamic features at different spatiotemporal resolutions. Through feature extraction and
merging mechanisms, it effectively integrates local video details with global semantic information.
By combining video features with the guiding semantic information of preset prompts to evaluate
videos, it can improve the match between the evaluation result and a target scene, thereby enhancing
25 the robustness and adaptability of video quality evaluation tasks.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] In order to more clearly illustrate the technical solutions in the embodiments of the present
application, the following briefly introduces the drawings required for use in the description of the
embodiments. Obviously, the drawings described below are only some embodiments of the present

application. For ordinary technicians in this field, other drawings can be obtained based on these drawings without any creative work.

[0016] FIG. 1 is a schematic structural diagram of an electronic device provided in an embodiment of the present application.

5 **[0017]** FIG. 2 is a first flow chart of a video evaluation method according to an embodiment of the present application.

[0018] FIG. 3 is a second flow chart of a video evaluation method provided in an embodiment of the present application.

10 **[0019]** FIG. 4 is a block diagram of a video evaluation apparatus according to an embodiment of the present application.

DETAILED DESCRIPTION

[0020] In order to make objectives, technical solutions and advantages of the present application clearer, the present application is described in detail below with reference to the accompanying drawings and specific embodiments.

15 **[0021]** It should be noted that, in the present application, "at least one" refers to one or more, and "more than one" refers to two or more. "And/or" describes the relationship between associated objects, indicating that three possible relationships exist. For example, "A and/or B" can mean: A alone, A and B together, and B alone, where A and B can be singular or plural. The terms "first," "second," "third," "fourth," and so on (if any) in the specification, claims, and drawings of the present application are
20 used to distinguish similar objects, not to describe a specific order or precedence.

[0022] In the embodiments of the present application, words such as "exemplary" or "for example" are used to indicate examples, illustrations, or descriptions. Any embodiment or design described as "exemplary" or "for example" in the embodiments of the present application should not be interpreted as being more preferred or more advantageous than other embodiments or designs. Specifically, the
25 use of words such as "exemplary" or "for example" is intended to present related concepts in a concrete manner. The following embodiments and features in the embodiments may be combined with each other unless there is a conflict.

[0023] With the development of artificial intelligence-generated content (AIGC) technology, AIGC videos are becoming increasingly widely used in e-commerce. For example, related video generation

models can automatically generate AIGC videos based on text prompts, such as product display videos, virtual model fitting demonstrations, and personalized advertising clips. AIGC technology significantly reduces video production costs, improves operational efficiency in e-commerce scenarios, and promotes an evolution of content production models towards intelligentization.

5 However, the quality of AIGC videos generated by AIGC technology varies widely.

[0024] Unlike an evaluation criteria for AIGC videos in other scenarios, e-commerce typically has relatively high quality requirements for AIGC videos. For example, the quality of AIGC videos in e-commerce scenarios depends not only on traditional visual metrics such as clarity and color, but more importantly, on specific dimensions such as motion smoothness (e.g., a smooth movement of
10 clothing), scene logic plausibility (e.g., whether virtual model's movements match product's function), object shape authenticity, and visual consistency between text prompts and presented content. Problems such as jitter, distortion, or choppy movement in AIGC videos can directly impact consumer trust and purchase intention.

[0025] Related AIGC video evaluation methods lack an understanding of the hierarchical dynamic
15 features of videos and have limited generalization capabilities, making it difficult to obtain efficient and accurate AIGC video evaluation results in e-commerce scenarios. For example, related AIGC video evaluation methods typically rely solely on content features of a single frame, ignoring a temporal correlation between video frames. Another example is that the related AIGC video evaluation methods have limited semantic understanding capabilities and struggle to capture deep
20 semantic information of video content (such as a logical rationality of product function demonstrations). Furthermore, related AIGC video evaluation methods have low generalization performance, making them difficult to adapt to e-commerce scenarios.

[0026] In order to solve the above problems, the video evaluation method provided by the present application can obtain a plurality of sub-videos based on a plurality of sampling rate, thereby
25 facilitating capturing of dynamic features at different spatiotemporal resolutions. Through a feature extraction and merging mechanism, it effectively integrates local video details with global semantic information. By combining video features with guiding semantic information of preset prompts to evaluate videos, a matching degree between an evaluation result and a target scene can be improved, thereby enhancing robustness and adaptability of tasks of evaluating video quality.

[0027] FIG. 1 is a schematic diagram of a structure of an electronic device provided in one embodiment of the present application. As shown in FIG. 1, a video evaluation method provided in this embodiment of the present application can be applied to an electronic device 10. The electronic device 10 can be a mobile phone, a tablet computer, a smart wearable device, an augmented reality (AR) device or a virtual reality (VR) device, a laptop computer, a netbook, an energy storage device, a power distribution device, or other electronic devices. This embodiment of the present application does not impose any restrictions on a specific type of the electronic device.

[0028] As shown in FIG. 1, the electronic device 10 may include a communication module 101, a storage device 102, a processor 103, an input/output (I/O) interface 104, and a bus 105. The processor 103 is coupled to the communication module 101, the storage device 102, and the I/O interface 104 via the bus 105.

[0029] The communication module 101 may include a wired communication module and/or a wireless communication module. The wired communication module may provide one or more wired communication solutions such as universal serial bus (USB) and controller area network (CAN). The wireless communication module may provide one or more wireless communication solutions such as wireless fidelity (Wi-Fi), Bluetooth (BT), mobile communication network, frequency modulation (FM), near field communication (NFC), infrared technology (IR), etc.

[0030] The storage device 102 may include one or more random access memories (RAMs) and one or more non-volatile memories (NVMs). The RAM can be directly read and written by the processor 103 and can be used to store executable programs (e.g., machine instructions) for an operating system or other running programs, as well as user data and application data. The RAM may include a static random-access memory (SRAM), a dynamic random access memory (DRAM), a synchronous dynamic random access memory (SDRAM), a double data rate synchronous dynamic random access memory (DDR SDRAM), etc.

[0031] The non-volatile storage device can also store executable programs, user data, and application data, etc., and can be pre-loaded into a random access memory for direct reading and writing by the processor 103. The non-volatile storage device can include a disk storage device and flash memory.

[0032] The storage device 102 is configured to store one or more computer programs. The one or more computer programs are configured to be executed by the processor 103. The one or more computer programs include a plurality of instructions. When the plurality of instructions are executed by the processor 103, the video evaluation method executed on the electronic device 10 can be implemented.

[0033] In other embodiments, the electronic device 10 further includes an external storage device interface for connecting to an external storage device to expand a storage capacity of the electronic device 10.

[0034] The processor 103 may include one or more processing units. For example, the processor 103 may include an application processor (AP), a modem processor, a graphics processing unit (GPU), an image signal processor (ISP), a controller, a video codec, a digital signal processor (DSP), a baseband processor, and/or a neural-network processing unit (NPU). Different processing units may be independent devices or integrated into one or more processors.

[0035] The processor 103 provides computing and control capabilities. For example, the processor 103 is configured to execute a computer program stored in the storage device 102 to implement the above-mentioned video evaluation method.

[0036] The I/O interface 104 provides a channel for user input and output. For example, the I/O interface 104 can be used to connect various input and output devices, such as a mouse, a keyboard, a touch screen device, and a display screen, allowing users to enter information or visualize information. Furthermore, the I/O interface 104 can also be used to connect various sensor devices, such as radar sensors and image sensors, to obtain required radar point cloud data, image data, and other data.

[0037] The bus 105 is at least used to provide a channel for mutual communication among the communication module 101, the storage device 102, the processor 103, and the I/O interface 104 in the electronic device 10.

[0038] It should be understood that the structure illustrated in the embodiments of the present application do not constitute a specific limitation on the electronic device 10. In other embodiments of the present application, the electronic device 10 may include more or fewer components than shown, or may combine or separate certain components, or arrange the components differently. The

illustrated components may be implemented in hardware, software, or a combination of software and hardware.

[0039] The following description will be based on an example of a computer program product implementing the video evaluation method of an embodiment of the present application running on an electronic device (e.g., the electronic device 10 shown in FIG. 1). FIG. 2 shows a flow chart of a video evaluation method provided in an embodiment of the present application. In one embodiment of the present application, the method includes the following steps:

[0040] Step S201, a video is sampled based on multiple sampling rates to obtain multiple sub-videos of the video.

[0041] In one embodiment of the present application, the video is an AIGC video generated based on artificial intelligence generated content (AIGC) technology, and a process of generating the AIGC video can be driven by a prompt engineering.

[0042] In one example, in an e-commerce scenario, an AIGC video generation model can receive a structured text instruction input by a user, such as a first prompt, and automatically generate a video showcasing a product based on the first prompt. The first prompt may include, but is not limited to, following text information: a core selling point of the product (e.g., "portable and foldable design, 20,000mAh battery life"), target audience characteristics (e.g., "young outdoor enthusiasts"), visual style parameters (e.g., "cyberpunk neon lighting effects"), and scenario-specific requirements (e.g., "15-second short video, 4K resolution, 60 frames per second (fps), including a tutorial clip").

[0043] The AIGC video generation model utilizes a diffusion-based video generation architecture, combined with a Transformer spatiotemporal encoder and a dynamic motion prediction module. By learning from a distribution of massive e-commerce video data, it transforms the first prompt into dynamic video content with commercial appeal. For example, if the first prompt instructs the user to "generate a virtual catwalk video of summer dresses," the AIGC video generation model can automatically synthesize dynamic fabric physics effects, multi-angle product displays, and background music that matches fashion trends, and output the AIGC video corresponding to the first prompt.

[0044] In one embodiment of the present application, the sampling of the video based on multiple sampling rates to obtain multiple sub-videos of the video, includes: determining multiple sampling

rates based on a frame rate of the video; for each sampling rate of the multiple sampling rates, extracting video frames from the video based on each sampling rate to obtain a video frame sequence corresponding to each sampling rate; and obtaining a sub-video corresponding to each sampling rate based on the video frame sequence corresponding to each sampling rate.

5 **[0045]** In one example, the multiple sampling rates can be determined based on the frame rate of the video. For example, a number of the multiple sampling rates can be proportional to the frame rate of the video. The sampling rate can indicate a number of video frames extracted from the video per second, and a sampling interval satisfies following formula: Sampling interval = (Frame rate of the video / Sampling Rate) - 1.

10 **[0046]** For example, if the frame rate of the video is 60fps, four sampling rates can be determined: 20fps, 15fps, 12fps, and 10fps. A sampling rate of 20fps means sampling 20 frames from 60 frames per second of the video, for example, a value of the sampling interval equals 2. A sampling rate of 15 fps means sampling 15 frames from the 60 frames per second of the video, for example, the value of the sampling interval equals 3. A sampling rate of 12 fps means sampling 12 frames from the 60
15 frames per second of the video, for example, the value of the sampling interval equals 4. A sampling rate of 10fps means sampling 10 frames from the 60 frames per second of the video, for example, the value of the sampling interval equals 5.

[0047] For example, if the frame rate of the video is 30 fps, two sampling rates can be determined: 15fps and 10fps. A sampling rate of 15fps means sampling 15 frames from the video's 30 frames per
20 second, for example, the value of the sampling interval equals 1. A sampling rate of 10fps means sampling 10 frames from the video's 30 frames per second, for example, the value of the sampling interval equals 2.

[0048] In one example, after extracting video frames from the video based on any one sampling rate of the multiple sampling rates, a video frame sequence corresponding to the any one sampling
25 rate can be constructed according to an order of the extracted video frames in the video. For example, if the frame rate of the video is 30fps and the sampling rate is 15fps, the video frames extracted from the video according to the sampling interval 1 can be constructed into a video frame sequence L1. If the frame rate of the video is 30fps and the sampling rate is 10 fps, the video frames extracted from the video according to the sampling interval 2 can be constructed into a video frame sequence L2. In

this way, the video frame sequence corresponding to each sampling rate of the multiple sampling rates can be obtained, thereby multiple video frame sequences are obtained.

[0049] In one example, a sub-video corresponding to any one sampling rate of the multiple sampling rates can be obtained based on the video frame sequence corresponding to the any one sampling rate. For example, continuing with the above example, the video with the frame rate of 30fps corresponds to two sampling rates, where the video frame sequence corresponding to the sampling rate of 15fps is L1, and the video frame sequence corresponding to the sampling rate of 10fps is L2. Based on the video frame sequence L1, a sub-video S1 corresponding to the sampling rate of 15fps can be obtained, and the frame rate of the sub-video S1 is 15fps. Based on the video frame sequence L2, a sub-video S2 corresponding to the sampling rate of 10fps can be obtained, and the frame rate of the sub-video S2 is 10fps. In this way, multiple sub-videos of the video (i.e., an original video) with different frame rates can be obtained based on the multiple video frame sequences.

[0050] Based on the above embodiments, dynamic sampling rate processing technology can be used to achieve multi-level sampling of AIGC-generated videos. By sampling the video at multiple frequencies and variable intervals at different temporal resolutions, multiple sub-videos with different temporal granularities can be constructed. This allows the capture of local motion smoothness and global motion coherence based on the sub-videos, effectively identifying timing anomalies such as flicker, jitter, and deformation, and significantly enhancing the model's ability to perceive motion smoothness.

[0051] Step S202, a first feature vector corresponding to each sub-video is obtained by extracting features from each sub-video using the video evaluation model, such that first feature vectors corresponding to all of the plurality of sub-videos are obtained.

[0052] In one embodiment of the present application, before using the video evaluation model to extract features from each sub-video, the video frames of each sub-video can be preprocessed. In one example, the preprocessing includes scaling the video frame. For example, each video frame can be scaled to a resolution of 448×448 or a resolution of 224×224 , etc., to meet a requirement of an input size of the video evaluation model and eliminate an impact of a resolution difference on feature extraction. In another example, the preprocessing can also include normalizing the video frame. For example, each pixel value of the video frame is scaled to an interval $[-1, 1]$ to map a distribution of

the pixel values of the video frame to an input domain during model training, thereby accelerating a convergence of a neural network.

[0053] In one embodiment of the present application, the video evaluation model includes a visual encoder, which can be used to extract features from each sub-video to obtain the first feature vector corresponding to each sub-video. For example, the visual encoder can be used to extract image features from each video frame of each sub-video to obtain the feature vector corresponding to each video frame. The feature vectors corresponding to all video frames of each sub-video are merged to obtain the first feature vector corresponding to each sub-video.

[0054] In one example, the visual encoder can be an Intern Vision Transformer (Intern ViT) model, a ViT model, or other models. The Intern ViT model is an optimized ViT model for multimodal applications. The following examples use the ViT model as the visual encoder to illustrate feature extraction for each sub-video.

[0055] In one example, when the ViT model extracts features from each video frame that has been input, the ViT model can segment the video frame into image patches and obtain an image patch vector corresponding to each image patch. A linear projection is performed on the image patch vector, and positional and temporal encodings are added to the image patch vector to obtain an embedding vector. The ViT model's attention head then performs calculation on the embedding vector to obtain the feature vector corresponding to each video frame.

[0056] For example, the ViT model can segment a video frame into multiple image patches corresponding to the video frame. For example, a video frame with a size $H \times W \times C$ can be evenly divided into multiple image patches of a same resolution, where $H \times W$ represents a resolution of the video frame and C represents a number of channels in the video frame. A resolution of the image patch can be, for example, 16×16 pixels, and a size of the image patch can be $16 \times 16 \times 3$.

[0057] When obtaining the image patch vector corresponding to the image patch, the ViT model can flatten each image patch into a vector to obtain the image patch vector. For example, an image patch of size $16 \times 16 \times 3$ can be flattened into a 768-dimensional vector.

[0058] The ViT model flattens and maps each image patch into a D -dimensional linear space using a learnable linear transformation matrix, generates an initial embedding vector for each image patch. D represents a preset integer that can be set as needed, for example, $D = 768$. A learnable position

encoding vector and a temporal encoding vector are added to each initial embedding vector to generate an embedding vector that includes both spatial position information and temporal position relationships.

[0059] The ViT model uses the Transformer encoder's multi-head attention mechanism to calculate the feature vector using the multiple attention heads on the embedding vectors. For example, a sequence of all embedding vectors corresponding to all image patches can be constructed as an input vector and fed into a deep network consisting of L stacked transformer encoder layers, where L represents a preset integer that can be set as needed, for example, $L = 12$. The input vector sequence can be passed to the k (for example, $k = 12$) attention heads of the encoder at the same time, and the k attention heads perform scaled dot product attention parallel calculation in parallel to obtain the first feature vector.

[0060] Based on the above embodiment, each attention head calculates a correlation (e.g., an attention weight) between each embedding vector and all other embedding vectors, thereby allowing the ViT model to simultaneously focus on the semantic information of different positions of the video frame and obtain new feature representations, such as the feature vector corresponding to each video frame.

[0061] In one example, when using the Intern ViT model for video frame feature extraction, its core process is optimized for multimodal scenarios based on the basic architecture of the ViT model.

[0062] For example, during an image segmentation stage, the Intern ViT model is no longer limited to uniform segmentation of video frames at a fixed resolution. Instead, it introduces dynamic segmentation and pixel shuffle technology to segment video frames. A size of the image patch can be dynamically adjusted based on the resolution of the input video frame (for example, 16×16 pixels or 32×32 pixels). Pixel shuffle is then used to downsample and reconstruct the segmented images, significantly reducing a number of visual tokens (for example, merging 4×4 tokens into 1). This avoids image distortion caused by traditional resizing operations when processing high-resolution (e.g., 4K) videos, while also improving computational efficiency.

[0063] During the feature mapping and encoding stages, the Intern ViT model continues the ViT model's flattening operation and the linear mapping mechanism. Each image patch is flattened into an image patch vector (for example, a $16 \times 16 \times 3$ image patch is flattened into a

768-dimensional vector). This image patch vector is mapped into a D-dimensional space (e.g., $D = 768$) using a learnable linear transformation matrix, yielding the initial embedding vector for each image patch.

[0064] However, unlike the ViT model, which adds positional and temporal encoding to the initial embedding vector of each image patch, the Intern ViT model uses variable-length visual position encoding (V2PE) technology. The V2PE technology can dynamically generate position information independent of an input resolution while implicitly incorporating temporal relationships, eliminating the need to add positional and temporal encoding vectors to the initial embedding vector of each image patch. This design not only simplifies the process but also achieves compatibility with the temporal position encoding of subsequent language models (such as InternLM and Qwen), laying the foundation for multimodal interaction (such as image-text alignment).

[0065] During an attention computation and multimodal alignment phase, similar to the ViT model, the Intern ViT model's Transformer encoder can adopt an architecture with L stacked layers (e.g., $L=12$) and k parallel attention heads (e.g., $k=12$). However, compared to the ViT model, the Intern ViT model enhances cross-modal capabilities through a phased training strategy, including, for example, a multi-layer perceptron (MLP) warm-up, ViT incremental learning, and full-model fine-tuning. The Intern ViT model also uses an MLP projector to map visual features into the semantic space of a language model, directly supporting complex tasks such as visual question answering and document understanding.

[0066] Based on the above embodiments, the Intern ViT model, while retaining the core mechanism of the ViT model, significantly improves multimodal compatibility and computational efficiency through technologies such as dynamic blocking, V2PE encoding, and pixel shuffling, making it more suitable for video-language joint task scenarios.

[0067] In one example, the feature vectors corresponding to all video frames of each sub-video can be merged to obtain a first feature vector corresponding to each sub-video. For example, the feature vectors corresponding to all video frames of each sub-video can be merged based on the order of the video frames in the sub-video using the Temporal Average Pooling method to obtain the first feature vector corresponding to each sub-video. In this way, the first feature vector contains feature

representations of all video frames in each sub-video, such as shallow features such as texture, and the dimensions of the first feature vectors corresponding to different sub-videos are consistent.

[0068] Step S203, the first feature vectors corresponding to all sub-videos are merged to obtain a merged feature vector.

5 **[0069]** In one embodiment of the present application, the first feature vectors corresponding to all sub-videos can be merged according to a channel splicing technology to obtain the merged feature vector. In one example, if the video has three sub-videos, three first feature vectors corresponding to the three sub-videos of the video can be merged to obtain a merged feature vector.

[0070] For example, the first feature vectors of all sub-videos are dimensionally aligned and
10 verified. If there are inconsistent feature lengths due to a difference of sampling rates, a zero padding or a fully connected layer projection is used to unify the dimensions of all first feature vectors, for example, to 256 dimensions. A splicing operation is performed along a feature channel dimension of the first feature vector to merge multiple first feature vectors of multiple sub-videos into a higher-dimensional merged feature vector. For example, three 256-dimensional first feature vectors can be
15 spliced to form a 768-dimensional (3×256) merged feature vector. To eliminate the dimensional differences between the features of different sub-videos, L2 normalization can be performed after splicing to ensure that the merged feature vector has a uniform numerical distribution.

[0071] Based on the above examples, compared to linear merging methods like weighted averaging, the channel splicing technology can fully preserve feature information of each sub-video, avoid
20 information loss, and significantly improve the fusion effect of multi-sampling rate features. At the same time, its computational complexity is far lower than dynamic weighting schemes such as temporal attention pooling, meeting the efficiency requirements of video processing scenarios.

[0072] Step S204, features are extracted from the merged feature vector using a video evaluation model to obtain a second feature vector.

25 **[0073]** In one embodiment of the present application, the video evaluation model includes a multi-layer perceptron (MLP). Using the video evaluation model to extract features from the merged feature vector to obtain the second feature vector includes: performing a nonlinear transformation on the merged feature vector using the MLP and obtaining a nonlinear transformation result;

compressing a number of channels (i.e., a dimensionality) of the nonlinear transformation result to a predetermined dimensionality, and obtaining the second feature vector.

[0074] In one example, the MLP may include a first fully connected layer, which is used to expand a dimension of the merged feature vector to a higher dimension to enhance its feature representation capabilities. For example, the first fully connected layer may project a 768-dimensional merged feature vector into a 4096-dimensional latent space to obtain a higher-dimensional feature representation of the merged feature vector. The MLP includes a nonlinear activation function (such as a Gaussian Error Linear Unit (GELU) function) to perform a nonlinear transformation on a vector output by the first fully connected layer to obtain a nonlinear transformation result. The MLP includes a second fully connected layer, which is used to compress the dimensionality of the nonlinear transformation result to a preset dimensionality (such as 512) to obtain the second feature vector. The second feature vector may include semantic, temporal, visual, and aesthetic information of the video.

[0075] Based on the above embodiment, the nonlinear expression of the merged feature vector can be deepened through a multi-layer perceptron, effectively improving the video evaluation model's ability to capture complex semantics of the video, and providing high-quality feature input for subsequent tasks such as video evaluation.

[0076] Step S205, based on preset prompt information and the second feature vector, the video is evaluated using the video evaluation model to obtain an evaluation result of the video.

[0077] In one embodiment of the present application, the preset prompt information includes first prompt information corresponding to the video, and the first prompt information is used to indicate the generation of the artificial intelligence-generated (AI-generated) video. For example, the first prompt information can refer to a description of the embodiment in step S201.

[0078] In one embodiment of the present application, the preset prompt information includes second prompt information corresponding to the video, and the second prompt information is used to instruct the video evaluation model to perform a preset evaluation operation on the video. The evaluation operation includes one or more of the following:

[0079] Evaluate a degree of semantic alignment between the video and the first prompt. For example, the second prompt might read: "Check whether the video meets the following requirements:

360-degree rotation of sneakers; background is a virtual runway with a brand logo; a copy is identical to 'Summer Breathable Running Shoes'."

[0080] Evaluate a temporal quality of the video, which includes one or more combinations of jitter, flicker, and continuity between adjacent video frames. For example, the second prompt might read,
5 "Check a timing stability of the video: jitter in fast-motion scenes must be less than 0.15; light switching flicker frequency must be less than 10Hz; and the subject's motion trajectory must conform to physical laws."

[0081] Evaluate an image quality of the video frames, including a clarity level and /or a noise level. For example, the second prompt message may read: "Check an image quality of the video: jitter in
10 fast-motion scenes must be less than 0.15; light switching flicker frequency must be less than 10Hz; and the subject's motion trajectory must conform to physical laws."

[0082] Evaluate an aesthetic quality of the video frames, including a content stability level and/or a color harmony level. For example, the second prompt might read: "Check an image quality: Edge clarity > 4.0 at 4K resolution; Noise level $< 35\text{dB}$; Colors consistent with brand color scheme ($\Delta E < 8$). "
15

[0083] Based on the above embodiment, the preset prompt information includes the second prompt information designed specifically for the video evaluation model, and the second prompt information is used to guide the video evaluation model to perform multi-dimensional automated evaluation operations on the video. Specific evaluation dimensions include: evaluation of a degree of semantic
20 alignment, which requires the video evaluation model to detect a degree of match between the content of the video and the text description given in the first prompt information through preset instructions; evaluation of a temporal quality, which focuses on a dynamic smoothness of the video, including indicators such as a jitter degree, a flicker degree, and a coherence degree; evaluation of an image quality, which involves basic image quality parameters such as a resolution and a noise control,
25 combining objective measurement with aesthetic standards. These structured second prompt information enable the video evaluation model to systematically output multi-dimensional evaluation results including semantic matching, dynamic smoothness, technical image quality, and visual aesthetics.

[0084] In one embodiment of the present application, the video evaluation model includes a large language model (LLM), such as InternLM2 or Qwen2. Based on preset prompt information and the second feature vector, the video is evaluated using the video evaluation model to obtain the video evaluation result, which includes: based on the second prompt information, using a pre-trained large language model to perform any evaluation operation on the video to obtain a corresponding evaluation result.

[0085] In one example, the first prompt information, the second prompt information, and the second feature vector can be used as input data and the input data are structured. For example, the second feature vector (e.g., visual features of the video), the first prompt information (e.g., instructions for generating the video), and the second prompt information (e.g., instructions for evaluating the video) can be encapsulated in a JSON structure to integrate the input data into an input format that the video evaluation model can understand.

[0086] Through multimodal fusion and encoding processing, a pre-trained cross-modal projector (e.g., an MLP layer) is used to map visual features into the semantic space of the LLM, achieving semantic alignment between visual information and textual instructions. Furthermore, the LLM is invoked to inject structured input data into the model's context window, and dynamic instruction injection techniques are used to guide the model in performing evaluation operations. For example, the model first understands a relevance of video content to generated instructions, and then performs multi-dimensional analysis of the video based on the evaluation instructions.

[0087] During model inference, the LLM generates text output including evaluation results, which can be parsed to extract structured metrics. For example, semantic alignment (e.g., whether the 360-degree display meets requirements) can be extracted through keyword matching or rule parsing, parameters of temporal quality (e.g., jitter value, flicker frequency) can be extracted through numerical value extraction, indicators of image quality (e.g., a resolution, a noise level) can be obtained through parameter parsing, and aesthetic quality scores (e.g., ΔE value for color harmony) can be converted through subjective evaluation.

[0088] The parsed multi-dimensional indicators are integrated into a JSON-formatted evaluation report, covering core dimensions such as semantic matching, dynamic smoothness, technical image quality, and visual aesthetics, to form a quantifiable video evaluation result. For example, the

evaluation result can be a score, such as a score range of [0, 5], where a higher score indicates a better evaluation result of the video.

[0089] Based on the above embodiment, by integrating multimodal data and dynamic instructions, LLM can simultaneously complete video content understanding and automatic evaluation, which is suitable for scenarios such as video quality monitoring and advertising material review.

[0090] In one embodiment of the present application, the method further includes: using a pre-training fine-tuning method and a preset loss function to train the video evaluation model until a preset loss function meets a preset condition.

[0091] In one example, a pre-trained large language model can be used as a base model, and a layered fine-tuning strategy can be employed to train the base model. For example, parameters of the base model's front layers (such as the Transformer encoder) can be frozen, and only later layers (such as the MLP layers) or newly added adaptation layers can be adjusted to preserve common knowledge and accelerate training. At the same time, a cross-modal projector can be used to align video feature vectors and textual prompt information into the same semantic space, achieving a joint visual-textual representation. A classification head or regression head can be added for video evaluation tasks, and the parameters of these modules can be fine-tuned.

[0092] In one example, the preset loss function may include, but is not limited to, a semantic alignment loss function, a temporal quality loss function, and a multi-task combination loss function. The semantic alignment loss function may be a cross-entropy loss function, used to evaluate the match between video content and generated instructions. The temporal quality loss function may be a mean square error loss function, used to quantify a deviation of indicators such as jitter and flicker from target values. The multi-task combination loss function may be a weighted sum of the semantic alignment loss function and the temporal quality loss function, used to balance optimization objectives across different evaluation dimensions.

[0093] In one example, the sample data for training the model includes sample videos, instructions for generating the sample videos, instructions for evaluating the sample videos, and true labels of the sample videos. For example, the true labels can be manually annotated evaluation results of the sample videos, such as ratings. The sample data can be divided into a training set and a validation set. The training set is used to train the model, and the validation set is used to determine the

corresponding loss value of a preset loss function. The preset loss function meeting a preset condition includes: the corresponding loss value being less than a preset loss threshold (e.g., 0.1).

[0094] Based on the above embodiment, by freezing most of parameters of the base model and pre-training and fine-tuning the base model, time and computing resource consumption for training the model can be significantly reduced. The multi-task combination loss function can simultaneously optimize multiple evaluation dimensions of the model, improving the model's generalization ability. By setting a preset loss threshold, overfitting of the training process of the model is avoided, ensuring the accuracy of the model's evaluation on unseen videos. This process enables the video evaluation model to efficiently adapt to specific task requirements, such as AIGC video evaluation in e-commerce scenarios, while maintaining the ability to understand and evaluate multimodal inputs.

[0095] The video evaluation method provided in the embodiments of the present application can generate multiple sub-videos based on multi-rate sampling, thereby facilitating the capture of dynamic features at different spatiotemporal resolutions. Through feature extraction and merging mechanisms, it effectively integrates local video details with global semantic information. By combining video features with the guiding semantic information of preset prompts to evaluate videos, it can improve the match between the evaluation results and the target scene, thereby enhancing the robustness and adaptability of video quality evaluation tasks.

[0096] FIG. 3 is a flow chart of a video evaluation method provided by another embodiment of the present application. The method includes: performing temporal pyramid sampling on a video, for example, performing temporal pyramid sampling on the video according to sampling rates 2fps, 3fps, and 4 fps separately, to obtain three different sub-videos. Each video frame in each sub-video is scaled to a preset size, for example, 448×448 pixels, to meet input requirements of the model. The three sub-videos are separately input into a VIT visual model (e.g., InternViT) to extract video features, resulting in three sets of video features, for example, a first feature vector corresponding to each sub-video. The video features corresponding to all sub-videos are merged (e.g., the video features corresponding to all sub-videos are concatenated), for example, to obtain a merged feature vector. The merged feature vector is input into a multi-layer perceptron (MLP) module to obtain features of a preset dimension, for example, a second feature vector with 512 dimensions. The second feature vector, the corresponding generation instruction (e.g., the first prompt) and the prompt word for the

dimension to be evaluated (e.g., the second prompt) are input into a large language model (LLM) to obtain a video evaluation result, such as a score.

[0097] Based on the above embodiment, efficient and accurate evaluation is achieved through multi-scale temporal sampling and cross-modal fusion technology: First, the video is sampled in a temporal pyramid to generate three sub-videos covering different time granularities to ensure that dynamic details are captured. Then, each sub-video frame is uniformly scaled to 448×448 pixels, and multi-scale video features are extracted using the visual model such as InternViT to obtain three sets of first feature vectors. After feature merging and dimensionality reduction using MLP projection, the second feature vector with 512 dimensions is formed. The second feature vector not only retains the temporal-spatial characteristics of the video but also adapts to the input requirements of the large language model. Finally, the second feature vector, generation instructions, and evaluation prompts are input into the LLM. Through cross-modal interaction, multi-dimensional evaluations such as semantic alignment, temporal quality, and image quality are achieved, and a structured scoring result is output. This method significantly improves the comprehensiveness and accuracy of video evaluation through multi-scale feature fusion and cross-modal interaction, while leveraging the efficiency of pre-trained models to achieve rapid adaptation to the needs of different evaluation tasks.

[0098] FIG. 4 is a structural diagram of a video evaluation apparatus provided in an embodiment of the present application.

[0099] In some embodiments, a video evaluation apparatus 40 may include multiple functional modules composed of computer program segments. The computer program for each segment in video evaluation apparatus 40 may be stored in a storage device of an electronic device and executed by at least one processor to perform video evaluation functions (see FIG. 2 for details).

[00100] In this embodiment, the video evaluation apparatus 40 can be divided into multiple functional modules based on the functions they perform. The functional modules may include: a video sampling module 401, a feature extraction module 402, a feature merging module 403, and a video evaluation module 404. As referred to herein, a module refers to a series of computer program segments that can be executed by at least one processor and can perform a fixed function, and is stored in a storage device. In this embodiment, the functional implementation of each module in the video

evaluation apparatus 40 can be found in the definition of the video evaluation method above, and will not be repeated here.

[00101] The video sampling module 401 is configured to sample a video based on multiple sampling rates to obtain multiple sub-videos of the video.

5 **[00102]** The feature extraction module 402 is configured to extract features from each sub-video using a video evaluation model to obtain a first feature vector corresponding to each sub-video.

[00103] The feature merging module 403 is configured to merge the first feature vectors corresponding to all sub-videos to obtain a merged feature vector.

10 **[00104]** The feature extraction module 402 is further configured to perform a feature extraction on the merged feature vector using the video evaluation model to obtain a second feature vector.

[00105] The video evaluation module 404 is configured to evaluate the video using the video evaluation model based on preset prompt information and the second feature vector to obtain an evaluation result of the video.

15 **[00106]** An embodiment of the present application also provides a computer-readable storage medium, on which a computer program is stored. The computer program includes program instructions. The method implemented when the program instructions are executed can refer to the methods in the above-mentioned embodiments of the present application.

20 **[00107]** The computer-readable storage medium may be an internal storage device of the electronic device of the above-mentioned embodiment, such as a hard disk or memory of the electronic device. The computer-readable storage medium may also be an external storage device of the electronic device, such as a plug-in hard disk equipped on the electronic device, a smart memory card (SMC), a secure digital (SD) card, a flash memory card, etc.

25 **[00108]** In some embodiments, the computer-readable storage medium may include a program storage area and a data storage area, wherein the program storage area may store an operating system, applications required for at least one function, etc.; the data storage area may store data created according to the use of the electronic device, etc.

[00109] In the above embodiments, the description of each embodiment has its own focus. For parts that are not described or recorded in detail in a certain embodiment, reference can be made to the relevant description of other embodiments.

[00110] Those skilled in the art will appreciate that the units and algorithm steps of each example described in conjunction with the embodiments disclosed herein can be implemented in electronic hardware, or a combination of computer software and electronic hardware. Whether these functions are performed in hardware or software depends on the specified application and design constraints of the technical solution. Professional and technical personnel can use different methods to implement the described functions for each specified application, but such implementation should not be considered beyond the scope of the present application.

[00111] In the embodiments provided in the present application, it should be understood that the disclosed devices/terminal devices and methods can be implemented in other ways. For example, the device/terminal device embodiments described above are merely illustrative. For example, the division of modules or units is merely a logical function division. In actual implementation, there may be other division methods, such as multiple units or components can be combined or integrated into another system, or some features can be ignored or not executed. Another point is that the mutual coupling or direct coupling or communication connection shown or discussed can be through some interfaces, indirect coupling or communication connection of devices or units, which can be electrical, mechanical or other forms.

[00112] Units described as separate components may or may not be physically separate, and components shown as units may or may not be physical units, that is, they may be located in one place or distributed across multiple network units. Some or all of these units may be selected to achieve the purpose of this embodiment according to actual needs.

[00113] The above embodiments are only used to illustrate the technical solutions of the present application, rather than to limit them. Although the present application has been described in detail with reference to the aforementioned embodiments, those skilled in the art should understand that they can still modify the technical solutions described in the aforementioned embodiments, or make equivalent replacements for some of the technical features therein. These modifications or replacements do not deviate the essence of the corresponding technical solutions from the spirit and scope of the technical solutions of the various embodiments of the present application, and should all be included in the scope of protection of the present application.

Claims

What is claimed is:

1. A video evaluation method, comprising:

sampling a video based on a plurality of sampling rates, and obtaining a plurality of sub-
5 videos of the video;

obtaining a first feature vector corresponding to each of the plurality of sub-videos by
extracting features from each of the plurality of sub-videos using a video evaluation model;

obtaining a merged feature vector by merging first feature vectors corresponding to all of
the plurality of sub-videos;

10 obtaining a second feature vector by extracting features from the merged feature vector
using the video evaluation model; and

obtaining an evaluation result of the video by evaluating the video using the video
evaluation model according to preset prompt information and the second feature vector.

2. The video evaluation method according to claim 1, wherein sampling the video based
15 on the plurality of sampling rates to obtain the plurality of sub-videos of the video comprises:

determining the plurality of sampling rates according to a frame rate of the video;

obtaining a video frame sequence corresponding to each of the plurality of sampling rates
by extracting video frames from the video according to each of the plurality of sampling rates;
and

20 obtaining a sub-video corresponding to each of the plurality of sampling rates according
to the video frame sequence corresponding to each of the plurality of sampling rates.

3. The video evaluation method according to claim 1, wherein the video evaluation model
comprises a visual encoder, and wherein obtaining the first feature vector corresponding to each
of the plurality of sub-videos by extracting features from each of the plurality of sub-videos
25 using the video evaluation model comprises:

obtaining a feature vector corresponding to each video frame of each of the plurality of
sub-videos by extracting image features from each video frame using the visual encoder; and

merging the feature vectors corresponding to all video frames of each of the plurality of sub-videos, and obtaining the first feature vector corresponding to each of the plurality of sub-videos.

4. The video evaluation method according to claim 1, wherein obtaining the merged feature vector by merging the first feature vectors corresponding to all of the plurality of sub-videos comprises:

obtaining the merged feature vector by merging the first feature vectors corresponding to all of the plurality of sub-videos according to a channel splicing technology.

5. The video evaluation method according to claim 1, wherein the video evaluation model comprises a multi-layer perceptron, and obtaining the second feature vector by extracting features from the merged feature vector using the video evaluation model comprises:

performing a nonlinear transformation on the merged feature vector using the multi-layer perceptron, and obtaining a nonlinear transformation result;

compressing a dimensionality of the nonlinear transformation result to a preset dimensionality, and obtaining the second feature vector.

6. The video evaluation method according to claim 1, wherein the preset prompt information comprises first prompt information corresponding to the video, the video is an artificial intelligence-generated video generated based on the first prompt information.

7. The video evaluation method according to any one of claims 1 to 6, wherein the preset prompt information comprises second prompt information corresponding to the video, and the second prompt information is used to instruct the video evaluation model to perform at least one preset evaluation operation on the video, and the at least one preset evaluation operation comprises one or more of the following:

evaluating a degree of semantic alignment between the video and the first prompt information;

evaluating a temporal quality of the video, where the temporal quality comprises a combination of one or more of a jitter degree, a flicker degree, and a coherence degree between adjacent video frames of the video;

evaluating an image quality of the video frames of the video, wherein the image quality comprises a clarity level and/or a noise level;

evaluating an aesthetic quality of the video frames of the video, wherein the aesthetic quality comprises a content stability level and/or a color harmony level.

5 8. The video evaluation method according to claim 7, wherein the video evaluation model comprises a large language model, and obtaining the evaluation result of the video by evaluating the video using the video evaluation model according to preset prompt information and the second feature vector comprises:

10 based on the second prompt information, performing any one of the evaluation operations on the video using the large language model, and obtaining a corresponding evaluation result.

9. The video evaluation method according to claim 1, further comprising:

training the video evaluation model using a pre-training fine-tuning method and a preset loss function until the preset loss function meets a preset condition.

10. An electronic device, comprising:

15 storage device, and

a processor, wherein the processor executes computer-readable instructions stored in the storage device to implement the video evaluation method according to any one of claims 1 to 9.

20

25