

METHOD OF INTERACTION BASED ON VIRTUAL CHARACTER, APPARATUS, DEVICE, MEDIUM, AND PRODUCT

FIELD

- 5 [0001] The disclosure relates to the technical field of artificial intelligence and, more particularly, to a method and apparatus of interaction based on a virtual character, a device, a medium, and a product.

BACKGROUND

- 10 [0002] At present, in video play scenarios such as movie watching, interaction between a user watching a video and a video content mainly relies on manual danmu (bullet comments) or social comments. Although some products embed AI assistants or other virtual characters, these virtual characters are activated only when the user explicitly calls on them: they may recommend titles, return search results, assist the user finding movie content of interest, or
15 impersonate a character from the show. All of these approaches are essentially reactive; none can initiate contact on its own, therefore the spontaneity and real-time engagement remain limited.

SUMMARY

- 20 [0003] The objects of embodiments of the disclosure are to provide a method and apparatus of interaction based on a virtual character, a device, a medium, and a product to solve the problem of poor experience of interaction based on a virtual character during video play. A specific technical solution is described as follows:
- [0004] In one aspect, the present disclosure provides a method of interaction based on a
25 virtual character, wherein the method is applied to a server device, and the method includes:
- [0005] when detecting video is played on a client device is detected, acquiring a target video emotion vector corresponding to a current play time point of the video from a video emotion vector sequence generated offline; wherein the video emotion vector sequence includes a plurality of video emotion vectors arranged according to play time points;
- 30 [0006] acquiring a user emotion vector collected in real time; and
- [0007] calling a preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and a prompt template for a video watching companion scenario, and sending the interaction information to the client device such that the interaction information is provided by a virtual character on the client device.

[0008] In a second aspect, the present disclosure provides an apparatus of interaction based on a virtual character, wherein the apparatus is applied to a server device, and the apparatus includes:

[0009] a target video emotion vector acquisition module configured to, when detecting
5 video is played on a client device, acquire a target video emotion vector corresponding to a current play time point of the video from a video emotion vector sequence generated offline; wherein the video emotion vector sequence includes a plurality of video emotion vectors arranged according to play time points;

[0010] a user emotion vector acquisition module configured to acquire a user emotion
10 vector collected in real time; and

[0011] an interaction information feedback module configured to call a preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and a prompt template for a video watching companion scenario, and send the interaction information to the client device such that the interaction information is
15 provided by a virtual character on the client device.

[0012] In a third aspect, the present disclosure provides an electronic device, including a processor, a memory, and computer programs stored in the memory and capable of running on the processor, wherein the computer programs, when executed by the processor, implement the above method.

[0013] In a fourth aspect, the present disclosure provides a computer-readable storage
20 medium, wherein the computer-readable storage medium has computer programs stored thereon, and the computer programs, when executed by a processor, implement the above method.

[0014] In a fifth aspect, the present disclosure provides a computer program product,
25 including computer programs, wherein the computer programs, when executed by a processor, implement the above method.

[0015] The embodiments of the present disclosure have the following beneficial effects.

[0016] In the embodiments of the present disclosure, when detecting the video is played on the client device is detected, the target video emotion vector corresponding to the current play
30 time point of the video is acquired from the video emotion vector sequence generated offline, then, the preset AI language model is called to generate the interaction information according to the target video emotion vector, the user emotion vector, and the prompt template for the video watching companion scenario, and the interaction information is sent to the client device such that the interaction information is provided by a virtual character on the client

device. Interaction with a user is achieved by the virtual character in conjunction with a video emotion generated offline and a user emotion collected in real time, eliminating the need for the user to take the initiative; emotion expressions that match the moment can thus be generated on the fly based on the video content and the user's emotional state, providing the user with a companion-style viewing solution, enhancing the proactivity and real-time performance of the interaction, and enhancing the immersive watching experience of the user.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] In order to describe the technical solutions in the present disclosure more clearly, accompanying drawings required to be used in the description for the disclosure will be briefly introduced below. Apparently, the accompanying drawings in the following description only show some embodiments of the present disclosure. Those skilled in the art can also derive other accompanying drawings from these accompanying drawings without creative work.

[0018] FIG. 1 is a schematic diagram of a system architecture provided in some embodiments of the present disclosure;

[0019] FIG. 2 is a flow chart showing a method of interaction based on a virtual character provided in some embodiments of the present disclosure; and

[0020] FIG. 3 is a schematic diagram of a timeline synchronization mechanism provided in some embodiments of the present disclosure;

[0021] FIG. 4 is a schematic diagram showing an apparatus of interaction based on a virtual character provided in some embodiments of the present disclosure;

[0022] FIG. 5 schematically shows a block diagram a computing processing device for performing a method of the present disclosure; and

[0023] FIG. 6 schematically shows a storage unit for holding or carrying a program code for implementing a method according to the present disclosure.

DETAILED DESCRIPTION

[0024] In order to make the above-mentioned objects, features and advantages of the present disclosure clearer and more understandable, the present disclosure will be further described in detail below in conjunction with the accompanying drawings and specific implementations. Apparently, the described embodiments are a part of embodiments of the present disclosure, not all the embodiments. Based on the embodiments in the present disclosure, all other embodiments obtained without creative work by those skilled in the art shall fall within the protective scope of the present disclosure.

[0025] In the embodiments of the present disclosure, at an offline stage, emotion recognition analysis is performed on multi-modal video information such as a picture, an audio, and a subtitle text in video to obtain a video emotion vector sequence, and at an online play stage of the video, multi-modal user information such as user emoji and voice is captured for emotion recognition analysis to obtain a user emotion vector. In conjunction with an AI language model, interaction information conforming to a plot tone, a character personality and a user state is automatically generated and is provided by a virtual character on a client device, such that emotional resonance interaction with an audience is achieved, problems of content understanding insufficiency, slow emotional response and single interaction form of a virtual character such as an AI assistant in the related art are solved, a companion watching solution is provided for a user, and immersive watching experience is enhanced.

[0026] As shown in FIG. 1, a video stream (i.e., video) is preprocessed to extract multi-modal video information such as a video frame, an audio frame and a subtitle text from the video stream, a visual model, an acoustic model and a semantic model are respectively adopted to perform visual emotion recognition, acoustic emotion recognition and semantic emotion recognition to obtain multi-modal emotion vectors such as a visual emotion vector, an acoustic emotion vector, and a semantic emotion vector, and multi-modal fusion is performed to obtain a video emotion vector.

[0027] During video play, a sensor such as a camera and a microphone is used to collect multi-modal user information such as user emoji features and user voice features, emotion recognition is performed on the multi-modal user information to obtain a user emotion vector, and a plot knowledge base is retrieved to obtain a plot text.

[0028] An event timeline is established according to play time points, video emotion vectors, user emotion vectors, and plot texts, these event stream messages are input into a large language model (LLM) as a video is played, and then, interaction information is generated and is provided by a virtual character on a client device, for example, danmu, voice, emoji, or other behaviors are output, such that an AI watching companion function is achieved.

[0029] Referring to FIG. 2, shown is a flow chart of a method of interaction based on a virtual character provided in some embodiments of the present disclosure. The method is applied to a server device which communicates with a client device. If the client device is a video play client device, the client device can provide an AI watching companion function for a user by a virtual character such as an AI assistant.

[0030] Specifically, the method may include the following steps:

[0031] Step 201, when detecting video is played on a client device, a target video emotion vector corresponding to a current play time point of the video is acquired from a video emotion vector sequence generated offline; wherein the video emotion vector sequence includes a plurality of video emotion vectors arranged according to play time points.

5 **[0032]** At an offline stage, emotion recognition can be performed on to-be-played video to generate video emotion vectors. The video emotion vectors carry corresponding play time points, so that the recognized video emotion vectors can be arranged according to the play time points to obtain the video emotion vector sequence for subsequent online play.

[0033] At an online play stage of the video, the video emotion vector sequence of the video
10 can be searched from a database, and according to the current play time point at which the client device plays back the video, the target video emotion vector matched with the current play time point can be accurately positioned and acquired from the video emotion vector sequence generated offline.

[0034] In some embodiments of the present disclosure, the step of when detecting video is
15 played on the client device, acquiring the target video emotion vector corresponding to the current play time point of the video from a video emotion vector sequence generated offline includes: when detecting the video is played on the client device, acquiring an event list; wherein the event list includes play time points corresponding to a plurality of marked events; and when the current play time point of the video matches a play time point of a target marked
20 event in the event list, acquiring the target video emotion vector corresponding to the current play time point from the video emotion vector sequence generated offline.

[0035] At the offline stage, specific events in the video can be recognized, and play time points corresponding to the specific events are marked to generate the event list. The plurality of marked events and the play time points corresponding to the marked events in the event list
25 are recorded. In some examples, by means of a timeline synchronization mechanism, as shown in FIG. 3, the marked events correspond to the play time points, such that play synchronization is achieved, and accurate triggering at corresponding time can be ensured.

[0036] At the online play stage of the video, the event list corresponding to the video can be acquired, and the current play time point of the video can be continuously monitored during
30 video play. Once the current play time point matches a play time point of a certain target marked event in the event list, the target video emotion vector corresponding to the current play time point can be accurately extracted from the video emotion vector sequence generated offline, which ensures that emotional features of a video at specific time can be quickly and accurately captured.

[0037] In some embodiments of the present disclosure, before the step of when detecting video is played on a client device, acquiring a target video emotion vector corresponding to the current play time point of the video from a video emotion vector sequence generated offline, the method further includes: acquiring the video, and for the plurality of play time
5 points of the video, performing multi-modal information extraction on the video to obtain multi-modal video information; calling a corresponding emotion recognition model according to an information type of the multi-modal video information to perform emotion recognition on the multi-modal video information to obtain first multi-modal emotion vectors; fusing the first multi-modal emotion vectors to obtain a first fused emotion vector; acquiring a basic
10 emotion vector corresponding to a plurality of basic emotion types, and determining first probability distribution corresponding to the plurality of basic emotion types according to the first fused emotion vector and the basic emotion vector; generating a video emotion vector at each play time point according to the first probability distribution; and arranging the video emotion vectors at the play time points to obtain the video emotion vector sequence.

15 [0038] At the offline stage, multi-modal information extraction can be performed on a picture, an audio and a subtitle text in the video to obtain multi-modal video information such as a video frame, an audio frame, and a subtitle text.

[0039] After the multi-modal video information is obtained, the emotion recognition models such as a visual model, an acoustic model and a semantic model can be called to
20 perform emotion recognition on the multi-modal video information to obtain the first multi-modal emotion vectors such as a visual emotion vector, an acoustic emotion vector, and a semantic emotion vector.

[0040] After the first multi-modal emotion vectors are obtained, multi-modal fusion can be performed on the first multi-modal emotion vectors to obtain an overall emotion vector, i.e.,
25 the first fused emotion vector.

[0041] For example, for visual emotion recognition: a video frame can be sampled from the video (for example, one frame is sampled per second), and then, a visual model is used for visual feature analysis to determine an overall emotion (such as tension, joy, sadness) of a picture so as to obtain a visual emotion vector, as shown in the following formula:

30 [0042] $V_t = f_v(I_t; \theta_v)$

[0043] wherein I_t is a video frame at time t , f_v is a visual model, θ_v is a model parameter, V_t is a visual emotion subvector, for example, the visual emotion subvector is a

512 to 1024-dimensional vector. In some examples, the visual model can be a convolutional neural network (CNN) model.

[0044] For example, for acoustic emotion recognition: an audio frame can be sampled from an audio stream of the video (such as subframe extraction), and then, an acoustic model can be used for acoustic feature analysis of features such as a tone, a speed, energy and a pitch to determine a speech emotion so as to obtain the acoustic emotion vector, as shown in the following formula:

$$[0045] \quad A_t = f_a(S_t; \theta_a)$$

[0046] wherein S_t is an audio frame of an audio signal window t , f_a is an acoustic model, θ_a is a model parameter, and A_t is an acoustic emotion subvector.

[0047] For example, for semantic emotion recognition: a subtitle text can be extracted from the video (or individual subtitle data), and then, a semantic model is used for analyzing semantic features to recognize a semantic emotion tone of a dialogue or a plot so as to obtain the semantic emotion vector, as shown in the following formula:

$$[0048] \quad T_t = f_t(W_t; \theta_t)$$

[0049] wherein W_t is a subtitle text at time t , f_t is a semantic model, θ_t is a model parameter, and T_t is a semantic emotion subvector.

[0050] After the multi-modal visual, audio and semantic emotion vectors V_t , A_t and T_t are obtained, multi-modal fusion can be performed to form an overall emotion vector so as to obtain a video emotion vector E_t . Specifically, the multi-modal emotion vectors can be fused by weighted summation, as shown in the following formula:

$$[0051] \quad E_t = w_v \cdot V_t + w_a \cdot A_t + w_t \cdot T_t$$

[0052] wherein w_v , w_a and w_t are weighting factors that can be set according to actual situations.

[0053] In the embodiments of the present disclosure, the multi-modal information extraction is performed on the video to obtain the multi-modal video information, and then, the corresponding emotion recognition models are called according to the information type of the multi-modal video information to perform emotion recognition on the multi-modal video information to obtain the first multi-modal emotion vectors, and the first multi-modal emotion vectors are fused to obtain the first fused emotion vector, such that comprehensive and accurate analysis for a video emotion is achieved.

[0054] In actual applications, the basic emotion vector corresponding to the plurality of basic emotion types, such as the basic emotion vector corresponding to five basic emotion types "joy, anger, sadness, fear, and surprise", can be preset.

[0055] After the first fused emotion vector which is a fixed-dimensional vector is obtained, the similarity between the first fused emotion vector and the basic emotion vector is calculated, and then, the first fused emotion vector is projected according to the similarity to obtain a probability value corresponding to the first fused emotion vector under each basic emotion type, i.e., the probability distribution, as shown in the following formula:

$$[0056] \quad P = \{P_{\text{joy}}, P_{\text{anger}}, P_{\text{sadness}}, P_{\text{fear}}, P_{\text{surprise}}\}$$

[0057] wherein P is the probability distribution, P_{joy} is a probability value corresponding to the basic emotion type "joy", P_{anger} is a probability value corresponding to the basic emotion type "anger", P_{sadness} is a probability value corresponding to the basic emotion type "sadness", P_{fear} is a probability value corresponding to the basic emotion type "fear", and P_{surprise} is a probability value corresponding to the basic emotion type "surprise".

[0058] For example, $P=(0.8, 0.1, 0.2, 0.3, 0.1)$ means that the probability value corresponding to the basic emotion type "joy" is 0.8, the probability value corresponding to the basic emotion type "anger" is 0.1, the probability value corresponding to the basic emotion type "sadness" is 0.2, the probability value corresponding to the basic emotion type "fear" is 0.3, and the probability value corresponding to the basic emotion type "surprise" is 0.1. In some examples, in the probability distribution, a basic emotion type corresponding to the maximum probability value is a global emotion label of the current video frame. For example, in $P=(0.8, 0.1, 0.2, 0.3, 0.1)$, 0.8 represents the probability value of the current "joy", and "joy" is the global emotion label of the current video frame.

[0059] After the first probability distribution is obtained, the video emotion vector at each play time point can be characterized in a form of probability distribution under a plurality of basic emotion types, and then, the video emotion vectors at the play time points can be arranged in chronological order to obtain the video emotion vector sequence for online use.

[0060] In the embodiments of the present disclosure, the basic emotion vector corresponding to the plurality of basic emotion types are acquired, the first probability distribution corresponding to the plurality of basic emotion types is determined according to the first fused emotion vector and the basic emotion vector, the video emotion vector at each play time point is generated according to the first probability distribution, and then, the video

emotion vectors at the play time points are arranged to obtain the video emotion vector sequence, such that accurate quantification and serialized storage of the video emotion are achieved.

[0061] In some embodiments of the present disclosure, after the step of arranging the video emotion vectors at the play time points to obtain the video emotion vector sequence, the method further includes: determining a plurality of target play time points from the video emotion vector sequence according to changes of the video emotion vectors in the video emotion vector sequence along the play time points; and marking events at the plurality of target play time points to obtain the event list.

[0062] Since the multi-modal video information such as the video frame, the audio frame and the subtitle text in video changes with play time, the extracted video emotion vectors also change with time, while the changes of the video emotion vectors reflect fluctuations and reversals of a video content plot. Therefore, the plurality of target play time points can be determined from the video emotion vector sequence according to the changes of the video emotion vectors, and then, the events can be marked at the plurality of target play time points to obtain the event list. When the subsequent online play reaches the target play time points corresponding to the marked events, an interaction function based on the video emotion vectors is triggered, such that key plots such as climaxes and reversals in a video are accurately captured, and interaction experience that is more in line with the development of the plot is provided for the user.

[0063] In some examples, the marked events may include various types of events such as high-energy events, reversal events, and calm events, wherein the high-energy events refer to plots in which emotions reach climaxes and are tense and exciting in the video, the reversal events refer to plots that suddenly change and are unexpected in the video, and the calm events refer to plots in which emotions are relatively stable and have no significant fluctuations in the video.

[0064] In some embodiments of the present disclosure, the step of determining the plurality of target play time points from the video emotion vector sequence according to changes of the video emotion vectors in the video emotion vector sequence along the play time points includes: for each video emotion vector in the video emotion vector sequence, determining a maximum probability value as an emotion intensity scalar from the first probability distribution corresponding to the plurality of basic emotion types, and arranging emotion intensity scalars at the play time points to obtain an emotion intensity scalar sequence; and determining target emotion intensity scalars from the emotion intensity scalar sequence, and

determining play time points corresponding to the target emotion intensity scalars as the target play time points; wherein the target emotion intensity scalars are larger than a preset scalar threshold, and are larger than emotion intensity scalars at previous and next play time points.

[0065] Since the high-energy events usually correspond to the plots in which the emotions reach the climaxes and are tense and exciting in the video, occurrence time of the high-energy events can be determined by finding play time points at which the probability values are larger than the threshold and are local maximum values in the first probability distribution of the video emotion vector sequence.

[0066] In actual applications, multi-dimensional video emotion vectors can be converted into emotion intensity scalars. Specifically, for each video emotion vector in the video emotion vector sequence, the maximum probability value can be determined as an emotion intensity scalar $S(t)$ from the first probability distribution corresponding to the plurality of basic emotion types, as shown in the following formula:

$$\text{[0067]} \quad S(t) = \max(P_{\text{joy}}, P_{\text{anger}}, P_{\text{sadness}}, P_{\text{fear}}, P_{\text{surprise}})$$

[0068] After the emotion intensity scalars are obtained, the emotion intensity scalars at the play time points can be arranged to obtain the emotion intensity scalar sequence, then, the emotion intensity scalars can be regarded as a curve for peak detection, the target emotion intensity scalars are found in the emotion intensity scalar sequence, and the play time points corresponding to the target emotion intensity scalars are determined as the target play time points corresponding to the high-energy events.

[0069] Specifically, the local maximum values in the emotion intensity scalar sequence are detected, that is, the emotion intensity scalars are larger than the emotion intensity scalars at the previous and next play time points, and the emotion intensity scalars larger than the preset scalar threshold are determined as the target emotion intensity scalars from the local maximum values, for example, the preset scalar threshold is 0.8.

[0070] In the embodiments of the present disclosure, the video emotion vectors are converted into the emotion intensity scalars, then, the target emotion intensity scalars larger than the preset scalar threshold and larger than the emotion intensity scalars at the previous and next play time points are determined from the emotion intensity scalar sequence, and the play time points corresponding to the target emotion intensity scalars are determined as the target play time points corresponding to the high-energy events, such that the occurrence time of the high-energy events in the video is accurately positioned, and the interaction experience is enhanced.

[0071] In some embodiments of the present disclosure, the step of determining the plurality of target play time points from the video emotion vector sequence according to changes of the video emotion vectors in the video emotion vector sequence along the play time points includes: determining a vector angle between video emotion vectors at adjacent play time points according to the first probability distribution of the video emotion vectors at the adjacent play time points in the video emotion vector sequence; and determining play time points at which the vector angle is larger than a preset angles the target play time points.

[0072] Since the reversal events usually occur at plot turning points and time when an emotion changes in a wider range, they can be determined by changes of angles between the video emotion vectors. Specifically, the video emotion vector is a multi-dimensional vector, the vector angle between the video emotion vectors at the adjacent play time points can be calculated according to the first probability distribution of the video emotion vectors at the adjacent play time points. If the vector angle is expressed as a cosine value, a formula is shown as follows:

$$\cos(\theta) = \frac{P(t) \cdot P(t-1)}{||P(t)|| ||P(t-1)||}$$

[0074] When the vector angle is larger than the preset angle, for example, the preset angle is 30° , the play time points can be determined as the target play time points corresponding to the reversal events in the video.

[0075] In the embodiments of the present disclosure, the vector angle between the video emotion vectors at the adjacent play time points is calculated, and the play time points at which a vector angle is larger than the preset angle are determined as the target play time points corresponding to the reversal events, such that occurrence time of the reversal events in the video is accurately positioned, and the interaction experience is enhanced.

[0076] Step 202: acquiring a user emotion vector collected in real time.

[0077] At the online play stage of the video, the user emotion vector can be collected in real time for analysis.

[0078] In some embodiments of the disclosure, the step of acquiring a user emotion vector collected in real time includes:

[0079] acquiring multi-modal user information collected in real time by a terminal device where the client device is installed; calling corresponding emotion recognition models according to an information type of the multi-modal user information to perform emotion recognition on the multi-modal user information to obtain second multi-modal emotion

vectors; fusing the second multi-modal emotion vectors to obtain a second fused emotion vector; acquiring a basic emotion vector corresponding to a plurality of basic emotion types, and determining second probability distribution corresponding to the plurality of basic emotion types according to the second fused emotion vector and the basic emotion vector;
5 and generating the user emotion vector according to the second probability distribution.

[0080] In some embodiments of the present disclosure, the multi-modal user information includes any one or more of the following information: user voice information, user facial information, and user input information.

[0081] In actual applications, the multi-modal user information can be collected in real time
10 by the terminal device where the client device is installed, for example, the multi-modal user information such as voice, input and facial expression is acquired by a sensor such as a camera and a microphone.

[0082] After the multi-modal user information is obtained, the emotion recognition models such as a visual model, an acoustic model and a semantic model can be called to perform
15 emotion recognition on the multi-modal user information to obtain the second multi-modal emotion vectors such as a visual emotion vector, an acoustic emotion vector, and a semantic emotion vector.

[0083] After the second multi-modal emotion vectors are obtained, multi-modal fusion can be performed on the second multi-modal emotion vectors to obtain an overall emotion vector,
20 i.e., the second fused emotion vector.

[0084] In the embodiments of the present disclosure, the multi-modal user information is collected in real time, then, the corresponding emotion recognition models are called according to the information type of the multi-modal user information to perform the emotion recognition on the multi-modal user information to obtain the second multi-modal emotion
25 vectors, and the second multi-modal emotion vectors are fused to obtain the second fused emotion vector, such that comprehensive and accurate analysis for a user emotion is achieved.

[0085] In actual applications, the basic emotion vector corresponding to the plurality of basic emotion types, such as the basic emotion vector corresponding to five basic emotion types "joy, anger, sadness, fear, and surprise", can be preset.

[0086] After the second fused emotion vector which is a fixed-dimensional vector is obtained, the similarity between the second fused emotion vector and the basic emotion vector is calculated, and then, the second fused emotion vector is projected according to the similarity to obtain a probability value corresponding to the second fused emotion vector
30

under each basic emotion type, i.e., the probability distribution, as shown in the following formula:

$$\text{[0087]} \quad P = \{P_{\text{joy}}, P_{\text{anger}}, P_{\text{sadness}}, P_{\text{fear}}, P_{\text{surprise}}\}$$

[0088] wherein P is the probability distribution, P_{joy} is a probability value corresponding to the basic emotion type "joy", P_{anger} is a probability value corresponding to the basic emotion type "anger", P_{sadness} is a probability value corresponding to the basic emotion type "sadness", P_{fear} is a probability value corresponding to the basic emotion type "fear", and P_{surprise} is a probability value corresponding to the basic emotion type "surprise".

10 **[0089]** For example, $p_{\text{user}} = (0.1, 0.1, 0.6, 0.2, 0.1)$ means that the probability value corresponding to the basic emotion type "joy" is 0.1, the probability value corresponding to the basic emotion type "anger" is 0.1, the probability value corresponding to the basic emotion type "sadness" is 0.6, the probability value corresponding to the basic emotion type "fear" is 0.2, and the probability value corresponding to the basic emotion type "surprise" is 0.1. In
15 some examples, in the probability distribution, a basic emotion type corresponding to the maximum probability value is the current global emotion label of a user. For example, in $P=(0.1, 0.1, 0.6, 0.2, 0.1)$, 0.6 represents the probability value of the current "sadness", and "sadness" is the current global emotion label of the user.

[0090] After the second probability distribution is obtained, the real-time user emotion
20 vector can be characterized in a form of probability distribution under a plurality of basic emotion types for online use, such that the user emotion is accurately quantified.

[0091] Step 203: calling a preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and a prompt template for a video watching companion scenario, and sending the interaction information to the client
25 device such that the interaction information is provided by a virtual character on the client device.

[0092] For example, the AI language model is the LLM.

[0093] In actual applications, the prompt template for the video watching companion scenario can be preset, and the prompt template can include various keywords, sentence
30 structures, or specific requirements for guiding the AI language model to generate the interaction information. After the target video emotion vector and the user emotion vector are obtained, the target video emotion vector, the user emotion vector and the selected prompt

template are input into the AI language model together, and according to the input information, the AI language model generates interaction information conforming to the current plot of the video and an emotional state of the user.

[0094] When the interaction information is generating, the target video emotion vector and the user emotion vector can be analyzed to understand current emotion atmosphere of the video and current emotion state of the user, and then, the interaction information is generating by selecting an adequate emoji code, a keyword or a sentence structure from a database, combining the prompt template for the video watching companion scenario, the current emotion atmosphere of the video and the current emotion state of the user.

10 **[0095]** For example, if the target video emotion vector displays that the video is currently at a high-energy, intense and exciting plot, while the user emotion vector displays that the current emotion of the user is relatively calm, it can be generated the interaction information such as “This plot is so intense and exciting—how come you look so calm? Have you already figured out what’s going to happen next?”. For another example, if the target video emotion
15 vector displays that the video plot is reversal, while the user emotion vector displays that current emotion of the user is relatively surprised, it can be generated the interaction information such as “Wow, this plot twist is completely unexpected—are you as surprised as I am?”.

[0096] After being generated, the interaction information can be sent to the client device.
20 After receiving the interaction information, the client device will provide the interaction information to the user by the preset virtual character in a form of voice, text, emoji or other forms. For example, the virtual character can say generated interaction sentences with vivid voice, or display the same in a form of dynamic text danmu on a video play interface, thereby enhancing the interaction experience between the user and the video, and making the user feel
25 a more thoughtful and interesting companionship when watching the video.

[0097] In specific application, after the client device receives the interaction information, the interaction information can include one or more types of information, such the interaction information as a voice text, a danmu text and an emoji code. For different types of the interaction information, the client device can process and output them in different ways.

30 **[0098]** For example, for the voice text, the client device can convert a receiving voice text as audio data, and then voice output is performed by the terminal device in the client device. In some examples, according to settings of the virtual character, the client device can adjust voice parameters such as pitch, speech rate and timbre, thereby making voice output better conform to personnel character personal traits of the virtual character.

[0099] For another example, for the danmu text, the client device can determine a displaying position of the danmu text on the video play interface, such as a danmu region, and then, display the danmu text in the danmu region by a form of a danmu message. In some examples, during the displaying process, the client device can select appropriate font color, size, and other text effects to display the danmu message according to emotion tendency of the danmu text.

[00100] For another example, for the emoji code, the client device can pre-store a plurality of emoji templates of the virtual character, such as joy, sadness and surprise. When the emoji code is received, the client device can determine a corresponding target emoji template from the preset emoji template, and then, output the target emoji template on the video play interface.

[00101] In some examples, the interaction information can be further optimized in conjunction with personal preferences and historical behavior data of the user to meet personalized needs of the user.

[00102] In some embodiments of the present disclosure, the step of calling a preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and the prompt template for the video watching companion scenario includes: determining a matching degree between the target video emotion vector and the user emotion vector; and when the matching degree is within a preset matching degree range, performing the calling a preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and a prompt template for a video watching companion scenario.

[00103] In some implementations, the matching degree between the target video emotion vector and the user emotion vector can be determined, for example, the matching degree is calculated by using a metric method such as cosine similarity and a Euclidean distance. When the calculated matching degree is within the preset matching degree range, it indicates that emotional states of the target video emotion vector and the user emotion vector are matched to a certain extent. At this time, the calling a preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and a prompt template for a video watching companion scenario is performed, which can make the generated interaction information more in line with the current user emotion and the plot development of the video.

[00104] In some embodiments of the present disclosure, the step of calling the preset AI language model is called to generate interaction information according to the target video

emotion vector, the user emotion vector, and the prompt template for the video watching companion scenario includes: determining a target plot text from a preset plot knowledge base according to the current play time point; acquiring the prompt template for the video watching companion scenario; filling the target video emotion vector, the user emotion vector and the target plot text in an input content prompt of the prompt template to obtain a prompt example; and calling the preset AI language model to generate the interaction information according to the prompt example.

[00105] The prompt template includes a system instruction prompt, an input content prompt, and a reply rule prompt; the system instruction prompt includes an identity prompt for indicating an identity of the virtual character as a video watching companion assistant and a target prompt for indicating that reply goals are to enhance interaction of a user and provide background knowledge; and the reply rule prompt includes a reply time point prompt for indicating a reply time point as a key explanation point of a plot, a reply tone prompt for indicating a reply tone, and a reply length limit.

[00106] In actual applications, the plot knowledge base can be retrieved according to the current play time point to obtain the target plot text, and then, in conjunction with the target video emotion vector, the user emotion vector, and the target plot text, the interaction information is generated. Not only are emotional changes in the video itself considered, but also real-time emotional feedback from the user and detailed information of the plot are infused, such that the interaction is more accurate and personalized.

[00107] In actual applications, the prompt template for the video watching companion scenario can be set, and the prompt template includes the system instruction prompt, the input content prompt, and the reply rule prompt. The obtained target video emotion vector, user emotion vector and target plot text are filled in the input content prompt of the prompt template to obtain the prompt example, and then, the prompt example is input into the AI language model to generate the interaction information.

[00108] The system instruction prompt includes the identity prompt for indicating the identity of the virtual character as the video watching companion assistant and the target prompt for indicating that the reply goals are to enhance interaction of the user and provide the background knowledge. For example, the identity prompt is "you are a humorous and knowledgeable watching companion assistant", and the target prompt is "your primary goal is to enhance interaction of a user and provide necessary background knowledge".

[00109] The reply rule prompt includes the reply time point prompt for indicating the reply time point as the key explanation point of the plot, the reply tone prompt for indicating the

reply tone, and the reply length limit. For example, the reply time point prompt is "if the current point is the key explanation point of the plot, please reply", the reply tone prompt is "the reply tone should be gentle and explanatory, which implies that this is a past content point", and the reply length limit is "reply length: no more than two sentences".

5 **[00110]** Exemplary description will be given below with an example:

[00111] For a structural example of a prompt:

[00112] [System instruction]: you are a humorous and knowledgeable watching companion assistant. Your primary goal is to enhance interaction of a user and provide necessary background knowledge.

10 **[00113]** [Current state input]:

[00114] User emotion vector: {

[00115] "joy": 0.8,

[00116] "anger": 0.1,

[00117] "sadness": 0.2,

15 **[00118]** "fear": 0.3,

[00119] "surprise": 0.1

[00120] }

[00121] Video emotion vector: {

[00122] "joy": 0.5,

20 **[00123]** "anger": 0.1,

[00124] "sadness": 0.2,

[00125] "fear": 0.1,

[00126] "surprise": 0.2

[00127] }

25 **[00128]** Video state: playing a key dialogue, keywords: [foreshadowing, memory]

[00129] Retrieved plot data (long-term memory): [the foreshadowing is about a necklace lost by a protagonist X in the first season, and it is associated with a motive of an action of a villain]

[00130] [Reply rules]:

30 **[00131]** 1. If the current point is the key explanation point of the plot, please reply.

[00132] 2. The reply tone should be gentle and explanatory, which implies that this is a past content point.

[00133] 3. Reply length: no more than two sentences.

[00134] [Output]: a reply is generated based on the above-mentioned information.

[00135] In some embodiments of the disclosure, the step of calling the preset AI language model to generate the interaction information according to the prompt example includes: acquiring reply format specification information; wherein the reply format specification information includes a plurality of interaction types and interaction content formats
5 corresponding to the interaction types; and calling the preset AI language model to generate the interaction information according to the prompt example and the reply format specification information.

[00136] In actual applications, the reply format specification information is preset, and the reply format specification information includes the plurality of interaction types and the
10 interaction content formats corresponding to the interaction types. When generating the interaction information according to the prompt example, the AI language model can generate interaction information conforming to a preset interaction type and a corresponding interaction content format.

[00137] The interaction information includes a target interaction type and an interaction
15 content conforming to an interaction content format corresponding to the target interaction type. The interaction types include any one of following types: voice interaction, danmu interaction, and emoji interaction; and an interaction content format corresponding to the voice interaction is a voice text, an interaction content format corresponding to the danmu interaction is a danmu text, and an interaction content format corresponding to the emoji
20 interaction is an emoji code. The emoji code can serve as an identification of the emoji template, different emoji code corresponding to different emoji template. The client device can pre-store the plurality of emoji templates of the virtual character. When the emoji code is received, the client device can determine a corresponding target emoji template from the preset emoji template, and then, outputting on the video play interface.

25 **[00138]** In some examples, the reply format specification information is shown as:

[00139] {
[00140] "action_type": "string", //action type: "voice", "danmu", "emoji"
[00141] "content": "string", //action content: voice text, danmu text, emoji code
[00142] }

30 **[00143]** For example, in a certain scenario, a watching companion generates a voice reply to obtain a result shown as follows:

[00144] {
[00145] "action_type": "voice",

[00146] "content": "Oh my god! This play is absolutely insane! That is literally god-like positioning!"

[00147] }

[00148] In some embodiments of the present disclosure, after the step of calling the preset
5 AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and the prompt template for the video watching companion scenario, the method further includes: determining a deviation between the target video emotion vector and the user emotion vector; and adjusting an emotional style of an interaction content in the interaction information and/or an interaction type of the interaction information
10 according to the deviation.

[00149] In actual applications, the deviation between the target video emotion vector and the user emotion vector can be determined, for example, a difference value in various emotional dimensions of the target video emotion vector and the user emotion vector is calculated. When
15 it is found that the deviation is greater, it means that there is a greater difference between the video emotion and the current user emotion. Therefore, the emotional style of the interaction content in the interaction information can be adjusted according to the deviation, and the interaction type of the interaction information can also be adjusted according to the deviation. By means of such adjustments, the interaction information can better conform to the current emotional state of the user, and the interaction experience between the user and the video is
20 further enhanced.

[00150] For example, if the video is relatively cheerful, but the user feels down, the emotional style of the interaction content can be adjusted to be more gentle and soothing so as to be better in line with the user emotion and avoid emotional impact. For example, when the deviation indicates that the user is relatively excited, the voice interaction type can be
25 selected for communicating with the user in a more direct and vivid way, such that an interaction effect is enhanced; and if the user is relatively calm, the danmu interaction type or the emoji interaction type can be adopted to provide the interaction information in a relatively relaxed way without interference with the user watching the video.

[00151] In some embodiments of the present disclosure, the step of adjusting an emotional
30 style of the interaction content in the interaction information and/or the interaction type of the interaction information according to the deviation includes:

[00152] in the case that basic emotion types corresponding to maximum probability values in the target video emotion vector and the user emotion vector are inconsistent, adjusting the emotional style of the interaction content in the interaction information to an emotional style

adapted to the basic emotion type corresponding to the maximum probability value in the user emotion vector, or adjusting the interaction type of the interaction information to an interaction type adapted to the basic emotion type corresponding to the maximum probability value in the user emotion vector.

5 **[00153]** In actual applications, when the basic emotion types corresponding to the maximum probability values in the target video emotion vector and the user emotion vector are inconsistent, it indicates that there is a significant difference between a video content emotion and the current user emotion, the user emotion is prioritized, and the emotional style and the interaction type in the interaction information are adjusted. According to these adjustment
10 strategies, it can be ensured that the generated interaction information is not only closely related to the video content, but also can be accurately in line with the current emotional state of the user, thereby providing more personalized and thoughtful video watching companion experience.

[00154] In some implementations, the basic emotion type such as "sadness" or "joy"
15 corresponding to the maximum probability value in the user emotion vector can be determined, and then, the tone, wording and the like in the interaction information are adjusted to a style adapted to this emotion type. For example, if the current user emotion is "sadness", a more comforting and understanding reply content can be generated to reflect resonance with the user emotion.

20 **[00155]** In other implementations, the interaction type of the interaction information can be adjusted to be better adapted to the current emotional state of the user. For example, when the user relatively feels down, a more gentle and non-intrusive interaction way such as the danmu interaction or the emoji interaction is selected to avoid emotional impact that may be caused by too direct or intense voice interaction. On the contrary, if the user is relatively positive or
25 excited, the voice interaction is selected for communicating with the user in a more direct and vivid way, such that the immediacy and fun of the interaction are enhanced.

[00156] In some embodiments of the disclosure, the step of adjusting an emotional style of the interaction content in the interaction information and/or the interaction type of the interaction information according to the deviation includes:

30 **[00157]** in the case that the basic emotion types corresponding to the maximum probability values in the target video emotion vector and the user emotion vector are consistent, and a probability value deviation between the maximum probability values is larger than a preset deviation threshold, determining a mean value of the maximum probability values in the target video emotion vector and the user emotion vector, and adjusting the emotional style of the

interaction content in the interaction information to an emotional style adapted to the mean value.

[00158] In actual applications, when the basic emotion types corresponding to the maximum probability values in the target video emotion vector and the user emotion vector are inconsistent, but the deviation between the maximum probability values exceeds the preset deviation threshold, for example, the preset deviation threshold is 0.5, it indicates that there is a significant difference between emotion intensities of the both, and therefore, by calculating the mean value of the maximum probability values in the target video emotion vector and the user emotion vector, the emotional style of the interaction content can be adjusted to a gentle state matched with the mean value. By means of such a dynamic balancing mechanism, not only can the consistency of the emotion types be maintained, but also more nuanced emotional adaptation can be achieved by intensity regulation.

[00159] For example, when the video and the user are both in the emotion type "joy", but the emotion intensity of the user is significantly higher than that of the video, a more lively interaction content can be generated; and if the video emotion is more intense, the interaction content will be appropriately restrained to avoid excessively stimulating the user.

[00160] In the embodiments of the present disclosure, when detecting the video is played on the client device is detected, the target video emotion vector corresponding to the current play time point of the video is acquired from the video emotion vector sequence generated offline, then, the preset AI language model is called to generate the interaction information according to the target video emotion vector, the user emotion vector, and the prompt template for the video watching companion scenario, and the interaction information is sent to the client device such that the interaction information is provided by a virtual character on the client device. Interaction with a user is achieved by the virtual character in conjunction with a video emotion generated offline and a user emotion collected in real time, eliminating the need for the user to take the initiative; emotion expressions that match the moment can thus be generated on the fly based on the video content and the user's emotional state, providing the user with a companion-style viewing solution, enhancing the proactivity and real-time performance of the interaction, and enhancing the immersive watching experience of the user.

[00161] It should be noted that, for the method embodiments, they are all described as a series of actions for the sake of brevity; however, those skilled in the art should understand that the embodiments of the present disclosure are not limited by the order of the described actions, because, in accordance with the embodiments of the present disclosure, some steps may be performed in other orders or simultaneously. In addition, those skilled in the art should

also understand that the embodiments described in the specification are preferred embodiments, and the actions involved are not necessarily essential to the embodiments of the present disclosure.

[00162] Referring to FIG. 4, shown is a schematic diagram of an apparatus of interaction
5 based on a virtual character provided in some embodiments of the present disclosure. The apparatus is applied to the server device. Specifically, the apparatus may include the following modules:

[00163] a target video emotion vector acquisition module 401 configured to, when detecting
10 video is played on a client device, acquire a target video emotion vector corresponding to a current play time point of the video from a video emotion vector sequence generated offline; wherein the video emotion vector sequence includes a plurality of video emotion vectors arranged according to play time points;

[00164] a user emotion vector acquisition module 402 configured to acquire a user emotion vector collected in real time; and

15 **[00165]** an interaction information feedback module 403 configured to call a preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and a prompt template for a video watching companion scenario, and send the interaction information to the client device such that the interaction information is provided by a virtual character on the client device.

20 **[00166]** In some embodiments of the present disclosure, the step of when detecting video is played on the client device, acquiring the target video emotion vector corresponding to the current play time point of the video from a video emotion vector sequence generated offline includes:

[00167] when detecting the video is played on the client device, acquiring an event list;
25 wherein the event list includes play time points corresponding to a plurality of marked events; and

[00168] when the current play time point of the video matches a play time point of a target marked event in the event list, acquiring the target video emotion vector corresponding to the current play time point from the video emotion vector sequence generated offline.

30 **[00169]** In some embodiments of the present disclosure, the step of calling the preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and the prompt template for the video watching companion scenario includes:

[00170] determining a matching degree between the target video emotion vector and the user emotion vector; and

[00171] when the matching degree is within a preset matching degree range, performing the calling a preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and a prompt template for a video watching companion scenario.

[00172] In some embodiments of the present disclosure, the step of calling the preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and the prompt template for the video watching companion scenario includes:

[00173] determining a target plot text from a preset plot knowledge base according to the current play time point;

[00174] acquiring the prompt template for the video watching companion scenario; wherein the prompt template includes a system instruction prompt, an input content prompt, and a reply rule prompt; the system instruction prompt includes an identity prompt for indicating an identity of the virtual character as a video watching companion assistant and a target prompt for indicating that reply goals are to enhance interaction of a user and provide background knowledge; and the reply rule prompt includes a reply time point prompt for indicating a reply time point as a key explanation point of a plot, a reply tone prompt for indicating a reply tone, and a reply length limit;

[00175] filling the target video emotion vector, the user emotion vector and the target plot text in the input content prompt of the prompt template to obtain a prompt example; and

[00176] calling the preset AI language model to generate the interaction information according to the prompt example.

[00177] In some embodiments of the present disclosure, the step of calling the preset AI language model to generate the interaction information according to the prompt example includes:

[00178] acquiring reply format specification information; wherein the reply format specification information includes a plurality of interaction types and interaction content formats corresponding to the interaction types; and

[00179] calling the preset AI language model to generate the interaction information according to the prompt example and the reply format specification information; wherein the interaction information includes a target interaction type and an interaction content conforming to an interaction content format corresponding to the target interaction type.

[00180] In some embodiments of the present disclosure, the interaction types include any one of following types: voice interaction, danmu interaction, and emoji interaction; and

[00181] an interaction content format corresponding to the voice interaction is a voice text, an interaction content format corresponding to the danmu interaction is a danmu text, and an

5 interaction content format corresponding to the emoji interaction is an emoji code.

[00182] In some embodiments of the present disclosure, the apparatus further includes:

[00183] a deviation determination module configured to determine a deviation between the target video emotion vector and the user emotion vector; and

[00184] an information adjustment module configured to adjust an emotional style of an

10 interaction content in the interaction information and/or an interaction type of the interaction information according to the deviation.

[00185] In some embodiments of the present disclosure, the step of adjusting the emotional style of the interaction content in the interaction information and/or the interaction type of the interaction information according to the deviation includes:

15 [00186] in the case that basic emotion types corresponding to maximum probability values in the target video emotion vector and the user emotion vector are inconsistent, adjusting the emotional style of the interaction content in the interaction information to an emotional style adapted to the basic emotion type corresponding to the maximum probability value in the user emotion vector, or adjusting the interaction type of the interaction information to an

20 interaction type adapted to the basic emotion type corresponding to the maximum probability value in the user emotion vector; and

[00187] in the case that the basic emotion types corresponding to the maximum probability values in the target video emotion vector and the user emotion vector are consistent, and a probability value deviation between the maximum probability values is larger than a preset

25 deviation threshold, determining a mean value of the maximum probability values in the target video emotion vector and the user emotion vector, and adjusting the emotional style of the interaction content in the interaction information to an emotional style adapted to the mean value.

[00188] In some embodiments of the present disclosure, the step of acquiring the user

30 emotion vector collected in real time includes:

[00189] acquiring multi-modal user information collected in real time by a terminal device where the client device is installed;

- [00190] calling corresponding emotion recognition models according to an information type of the multi-modal user information to perform emotion recognition on the multi-modal user information to obtain second multi-modal emotion vectors;
- [00191] fusing the second multi-modal emotion vectors to obtain a second fused emotion
5 vector;
- [00192] acquiring a basic emotion vector corresponding to a plurality of basic emotion types, and determining second probability distribution corresponding to the plurality of basic emotion types according to the second fused emotion vector and the basic emotion vector; and
- 10 [00193] generating the user emotion vector according to the second probability distribution.
- [00194] In some embodiments of the present disclosure, the multi-modal user information includes any one or more of the following information: user voice information, user facial information, and user input information.
- [00195] In some embodiments of the present disclosure, the apparatus further includes:
- 15 [00196] a multi-modal video information acquisition module configured to acquire video, and for a plurality of play time points of the video, perform multi-modal information extraction on the video to obtain multi-modal video information;
- [00197] a first multi-modal emotion vector acquisition module configured to call a corresponding emotion recognition model according to an information type of the multi-
20 modal video information to perform emotion recognition on the multi-modal video information to obtain first multi-modal emotion vectors;
- [00198] a first fused emotion vector fusion module configured to fuse the first multi-modal emotion vectors to obtain a first fused emotion vector;
- [00199] a first probability distribution determination module configured to acquire a basic
25 emotion vector corresponding to a plurality of basic emotion types, and determine first probability distribution corresponding to the plurality of basic emotion types according to the first fused emotion vector and the basic emotion vector;
- [00200] a video emotion vector generation module configured to generate a video emotion vector at each play time point according to the first probability distribution; and
- 30 [00201] a video emotion vector sequence acquisition module configured to arrange the video emotion vectors at the play time points to obtain the video emotion vector sequence.
- [00202] In some embodiments of the present disclosure, the apparatus further includes:

[00203] a target play time point determination module configured to determine a plurality of target play time points from the video emotion vector sequence according to changes of the video emotion vectors in the video emotion vector sequence along the play time points; and

[00204] an event list acquisition module configured to mark events at the plurality of target
5 play time points to obtain the event list.

[00205] In some embodiments of the present disclosure, the step of determining the plurality of target play time points from the video emotion vector sequence according to changes of the video emotion vectors in the video emotion vector sequence along the play time points includes:

10 [00206] for each video emotion vector in the video emotion vector sequence, determining a maximum probability value as an emotion intensity scalar from the first probability distribution corresponding to the plurality of basic emotion types, and arranging emotion intensity scalars at the play time points to obtain an emotion intensity scalar sequence; and

[00207] determining target emotion intensity scalars from the emotion intensity scalar
15 sequence, and determining play time points corresponding to the target emotion intensity scalars as the target play time points;

[00208] wherein the target emotion intensity scalars are larger than a preset scalar threshold, and are larger than emotion intensity scalars at previous and next play time points.

[00209] In some embodiments of the present disclosure, the step of determining the plurality
20 of target play time points from the video emotion vector sequence according to changes of the video emotion vectors in the video emotion vector sequence along the play time points includes:

[00210] determining a vector angle between video emotion vectors at adjacent play time points according to the first probability distribution of the video emotion vectors at the
25 adjacent play time points in the video emotion vector sequence; and

[00211] determining play time points at which the vector angle is larger than a preset angle as the target play time points.

[00212] In some embodiments of the present disclosure further provide an electronic device, including a processor, a memory, and a computer program stored on the memory and
30 executable on the processor, the processor executes the computer program, the processor implements the above method.

[00213] In some embodiments of the present disclosure further provide a computer-readable storage medium, having stored there on a computer program, when the computer program is executed by a processor, the computer program implements the above method.

[00214] In some embodiments of the present disclosure further provide a computer program product, which stores a computer program, the computer program implements the above method.

[00215] For apparatus embodiments, since they are substantially similar to method
5 embodiments, the description is relatively concise. For relevant details, reference may be made to the description of the method embodiments.

[00216] It should be noted that the user information (including but not limited to user device information, user personal information, etc.) and data (including but not limited to data used for analysis, stored data, displayed data, etc.) involved in the present disclosure are all
10 information and data authorized by users or fully authorized by all relevant parties. The collection, use, and processing of relevant data must comply with the applicable laws, regulations, and standards of the relevant countries and regions. Corresponding operational interfaces are provided to allow users to choose to authorize or deny such actions.

[00217] The various embodiments described in the present specification are presented in a
15 progressive manner. Each embodiment focuses on the differences from the others. For the parts that are the same or similar among the embodiments, reference to one another is sufficient.

[00218] Those skilled in the art should understand that embodiments of the present disclosure may be provided as methods, apparatus, or computer program products. Therefore,
20 the embodiments of the present disclosure may take the form of a complete hardware embodiment, a complete software embodiment, or an embodiment combining software and hardware aspects. Furthermore, the embodiments of the present disclosure may take the form of a computer program product implemented on one or more computer-usable storage medium (including, but not limited to, disk storage, CD-ROM, optical storage, etc.) including
25 computer-usable program code.

[00219] The embodiments of the present disclosure are described with reference to flowcharts and/or block diagrams of methods, terminal devices (systems), and computer program products according to the embodiments of the present disclosure. It should be understood that each process and/or block in the flowcharts and/or block diagrams, as well as
30 combinations of processes and/or blocks in the flowcharts and/or block diagrams, may be implemented by computer program instructions. These computer program instructions may be provided to a processor of a general-purpose computer, special-purpose computer, embedded processor, or other programmable data processing terminal device to produce a machine, so that the instructions executed by the processor of the computer or other

programmable data processing terminal device create means for implementing the functions specified in one or more processes of the flowchart and/or one or more blocks of the block diagram.

5 [00220] These computer program instructions may also be stored in a computer-readable memory capable of directing a computer or other programmable data processing terminal device to operate in a specific manner, so that the instructions stored in the computer-readable memory produce an article of manufacture including instruction means that implement the functions specified in one or more processes of the flowchart and/or one or more blocks of the block diagram.

10 [00221] These computer program instructions may also be loaded onto a computer or other programmable data processing terminal device, causing a series of operational steps to be performed on the computer or other programmable terminal device to produce computer-implemented processing, so that the instructions executed on the computer or other programmable terminal device provide steps for implementing the functions specified in one or more processes of the flowchart and/or one or more blocks of the block diagram.

[00222] The various component embodiments of the present disclosure may be implemented in hardware, or in software modules running on one or more processors, or in a combination thereof. It should be understood by those skilled in the art that a microprocessor or a digital signal processor (DSP) may be used in practice to realize some or all functions of some or all components in the computing processing device according to the embodiments of the present disclosure. The present disclosure may also be implemented as a device or system program (for example, a computer program and a computer program product) for performing part or all of the methods described herein. Such a program for implementing the present disclosure may be stored on a computer-readable medium, or may be in the form of one or more signals.

20

25 Such a signal may be downloaded from an Internet website, or provided on a carrier signal, or provided in any other form.

[00223] For example, FIG. 5 shows a block diagram of a computing processing device for performing the method according to the present disclosure. The computing processing device conventionally includes a processor 1010 and a computer program product or a computer-readable medium in the form of a memory 1020. The memory 1020 may be an electronic memory such as a flash memory, an electrically erasable programmable read-only memory (EEPROM), a EPROM, a hard disk, or a read only memory (ROM). The memory 1020 includes a storage space 1030 for program codes 1031 for executing any of the method steps in the above method. For example, the storage space 1030 for the program codes may include

30

each program code 1031 for respectively implementing various steps in the above method. These program codes may be read from or written into one or more computer program products. These computer program products include program code carriers such as hard disks, compact discs (CDs), memory cards or floppy disks. Such computer program products are
5 usually portable or fixed storage units as described with reference to FIG. 6. The storage unit may have memory segments and storage space arranged similarly to the memory 1020 in the computing processing device of FIG. 5. The program codes may, for example, be compressed in a suitable format. Typically, the storage unit includes computer-readable codes 1031', that is, codes that can be read by a processor such as 1010, these codes, when executed by the
10 computing processing device, cause the computing processing device to perform the various steps in the method described above.

[00224] Although preferred embodiments of the embodiments of the present disclosure have been described, once those skilled in the art have known the essential inventive concept, they may make further variations and modifications on these embodiments. Therefore, the
15 appended claims are intended to be interpreted as including the preferred embodiments and all of the variations and modifications that fall within the scope of the embodiments of the present disclosure.

[00225] Finally, it should be noted that in the present text, relational terms such as "first" and "second" are merely intended to distinguish one entity or operation from another entity
20 or operation, and that does not necessarily require or imply any actual relationship or order between such entities or operations. Moreover, the terms "include," "comprise," or any variations thereof are intended to cover non-exclusive inclusion, so that processes, methods, articles or terminal devices that include a series of elements do not only include these elements, but also include other elements that are not explicitly listed, or include elements that are
25 inherent to such processes, methods, articles or terminal devices. Unless further limitation is set forth, an element defined by the wording "including a ..." does not exclude additional same elements in the process, method, article, or terminal device including the element.

[00226] The method of interaction based on a virtual character, the apparatus, the device, the medium, and the product according to the present disclosure have been described in details
30 above. The principle and the embodiments of the present disclosure are described herein with reference to the particular examples, and the description of the above embodiments is merely intended to facilitate to understand the method according to the present disclosure and its core concept. Moreover, for those skilled in the art, according to the concept of the present disclosure, the particular embodiments and the range of applications may be varied. In

conclusion, the contents of the description should not be understood as limiting the present disclosure.

CLAIMS

1. A method of interaction based on a virtual character, wherein the method is applied to
5 a server device, and the method comprises:

when detecting video is played on a client device, acquiring a target video emotion vector corresponding to a current play time point of the video from a video emotion vector sequence generated offline; wherein the video emotion vector sequence comprises a plurality of video emotion vectors arranged according to play time points;

10 acquiring a user emotion vector collected in real time; and

calling a preset artificial intelligence (AI) language model to generate interaction information according to the target video emotion vector, the user emotion vector, and a prompt template for a video watching companion scenario, and sending the interaction information to the client device such that the interaction information is provided by a virtual
15 character on the client device.

2. The method according to claim 1, wherein the step of when detecting video is played on the client device, acquiring the target video emotion vector corresponding to the current play time point of the video from a video emotion vector sequence generated offline
20 comprises:

when detecting the video is played on the client device, acquiring an event list; wherein the event list comprises play time points corresponding to a plurality of marked events; and

when the current play time point of the video matches a play time point of a target marked event in the event list, acquiring the target video emotion vector corresponding to the current
25 play time point from the video emotion vector sequence generated offline.

3. The method according to claim 1, wherein the step of calling the preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and the prompt template for the video watching companion scenario
30 comprises:

determining a matching degree between the target video emotion vector and the user emotion vector; and

when the matching degree is within a preset matching degree range, performing the calling a preset AI language model to generate interaction information according to the target

video emotion vector, the user emotion vector, and a prompt template for a video watching companion scenario.

4. The method according to claim 1, wherein the step of calling the preset AI language
5 model to generate interaction information according to the target video emotion vector, the
user emotion vector, and the prompt template for the video watching companion scenario
comprises:

determining a target plot text from a preset plot knowledge base according to the current
play time point;

10 acquiring the prompt template for the video watching companion scenario; wherein the
prompt template comprises a system instruction prompt, an input content prompt, and a reply
rule prompt; the system instruction prompt comprises an identity prompt for indicating an
identity of the virtual character as a video watching companion assistant and a target prompt
for indicating that reply goals are to enhance interaction of a user and provide background
15 knowledge; and the reply rule prompt comprises a reply time point prompt for indicating a
reply time point as a key explanation point of a plot, a reply tone prompt for indicating a reply
tone, and a reply length limit;

filling the target video emotion vector, the user emotion vector and the target plot text in
the input content prompt of the prompt template to obtain a prompt example; and

20 calling the preset AI language model to generate the interaction information according
to the prompt example.

5. The method according to claim 4, wherein the step of calling the preset AI language
model to generate the interaction information according to the prompt example comprises:

25 acquiring reply format specification information; wherein the reply format specification
information comprises a plurality of interaction types and interaction content formats
corresponding to the interaction types; and

calling the preset AI language model to generate the interaction information according
to the prompt example and the reply format specification information; wherein the interaction
30 information comprises a target interaction type and an interaction content conforming to an
interaction content format corresponding to the target interaction type.

6. The method according to claim 5, wherein the interaction types comprise any one of
following types: voice interaction, danmu interaction, and emoji interaction; and

an interaction content format corresponding to the voice interaction is a voice text, an interaction content format corresponding to the danmu interaction is a danmu text, and an interaction content format corresponding to the emoji interaction is an emoji code.

5 7. The method according to claim 1, wherein after the step of calling the preset AI language model to generate interaction information according to the target video emotion vector, the user emotion vector, and the prompt template for the video watching companion scenario, the method further comprises:

10 determining a deviation between the target video emotion vector and the user emotion vector; and

adjusting an emotional style of an interaction content in the interaction information and/or an interaction type of the interaction information according to the deviation.

15 8. The method according to claim 7, wherein the step of adjusting the emotional style of the interaction content in the interaction information and/or the interaction type of the interaction information according to the deviation comprises:

20 in the case that basic emotion types corresponding to maximum probability values in the target video emotion vector and the user emotion vector are inconsistent, adjusting the emotional style of the interaction content in the interaction information to an emotional style adapted to the basic emotion type corresponding to the maximum probability value in the user emotion vector, or adjusting the interaction type of the interaction information to an interaction type adapted to the basic emotion type corresponding to the maximum probability value in the user emotion vector; and

25 in the case that the basic emotion types corresponding to the maximum probability values in the target video emotion vector and the user emotion vector are consistent, and a probability value deviation between the maximum probability values is larger than a preset deviation threshold, determining a mean value of the maximum probability values in the target video emotion vector and the user emotion vector, and adjusting the emotional style of the interaction content in the interaction information to an emotional style adapted to the mean
30 value.

9. The method according to claim 1, wherein the step of acquiring the user emotion vector collected in real time comprises:

acquiring multi-modal user information collected in real time by a terminal device where the client device is installed;

calling corresponding emotion recognition models according to an information type of the multi-modal user information to perform emotion recognition on the multi-modal user information to obtain second multi-modal emotion vectors;

fusing the second multi-modal emotion vectors to obtain a second fused emotion vector;

acquiring a basic emotion vector corresponding to a plurality of basic emotion types, and determining second probability distribution corresponding to the plurality of basic emotion types according to the second fused emotion vector and the basic emotion vector; and

generating the user emotion vector according to the second probability distribution.

10. The method according to claim 8, wherein the multi-modal user information comprises any one or more of the following information: user voice information, user facial information, and user input information.

11. The method according to any one of claims 1 to 10, wherein before the step of when detecting video is played on a client device, acquiring a target video emotion vector corresponding to the current play time point of the video from a video emotion vector sequence generated offline, the method further comprises:

acquiring video, and for a plurality of play time points of the video, performing multi-modal information extraction on the video to obtain multi-modal video information;

calling a corresponding emotion recognition model according to an information type of the multi-modal video information to perform emotion recognition on the multi-modal video information to obtain first multi-modal emotion vectors;

fusing the first multi-modal emotion vectors to obtain a first fused emotion vector;

acquiring a basic emotion vector corresponding to a plurality of basic emotion types, and determining first probability distribution corresponding to the plurality of basic emotion types according to the first fused emotion vector and the basic emotion vector;

generating a video emotion vector at each play time point according to the first probability distribution; and

arranging the video emotion vectors at the play time points to obtain the video emotion vector sequence.

12. The method according to claim 11, wherein after the step of arranging the video emotion vectors at the play time points to obtain the video emotion vector sequence, the method further comprises:

- determining a plurality of target play time points from the video emotion vector sequence
- 5 according to changes of the video emotion vectors in the video emotion vector sequence along the play time points; and
- marking events at the plurality of target play time points to obtain the event list.

13. The method according to claim 12, wherein the step of determining the plurality of
10 target play time points from the video emotion vector sequence according to changes of the video emotion vectors in the video emotion vector sequence along the play time points comprises:

- for each video emotion vector in the video emotion vector sequence, determining a maximum probability value as an emotion intensity scalar from the first probability
- 15 distribution corresponding to the plurality of basic emotion types, and arranging emotion intensity scalars at the play time points to obtain an emotion intensity scalar sequence; and
- determining target emotion intensity scalars from the emotion intensity scalar sequence, and determining play time points corresponding to the target emotion intensity scalars as the target play time points;
- 20 wherein the target emotion intensity scalars are larger than a preset scalar threshold, and are larger than emotion intensity scalars at previous and next play time points.

14. The method according to claim 12, wherein the step of determining the plurality of target play time points from the video emotion vector sequence according to changes of the
25 video emotion vectors in the video emotion vector sequence along the play time points comprises:

- determining a vector angle between video emotion vectors at adjacent play time points according to the first probability distribution of the video emotion vectors at the adjacent play time points in the video emotion vector sequence; and
- 30 determining play time points at which the vector angle is larger than a preset angle as the target play time points.

15. An apparatus of interaction based on a virtual character, wherein the apparatus is applied to a server device, and the apparatus comprises:

a target video emotion vector acquisition module configured to, when detecting video is played on a client device, acquire a target video emotion vector corresponding to a current play time point of the video from a video emotion vector sequence generated offline; wherein the video emotion vector sequence comprises a plurality of video emotion vectors arranged
5 according to play time points;

a user emotion vector acquisition module configured to acquire a user emotion vector collected in real time; and

an interaction information feedback module configured to call a preset AI language model to generate interaction information according to the target video emotion vector, the
10 user emotion vector, and a prompt template for a video watching companion scenario, and send the interaction information to the client device such that the interaction information is provided by a virtual character on the client device.

16. An electronic device, characterized by comprising a processor, a memory, and
15 computer programs stored in the memory and capable of running on the processor, wherein the computer programs, when executed by the processor, implement the method according to any one of claims 1 to 14.

17. A computer-readable storage medium, wherein the computer-readable storage
20 medium has computer programs stored thereon, and the computer programs, when executed by a processor, implement the method according to any one of claims 1 to 14.

18. A computer program product, comprising computer programs, wherein the computer programs, when executed by a processor, implement the method according to any one of
25 claims 1 to 14.