



(51) International Patent Classification:

G06F 16/738 (2019.01)

(21) International Application Number:

PCT/US2023/019039

(22) International Filing Date:

19 April 2023 (19.04.2023)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

202210656882.1 10 June 2022 (10.06.2022) CN

(71) Applicant: MICROSOFT TECHNOLOGY LICENSING, LLC [US/US]; One Microsoft Way, Redmond, Washington 98052-6399 (US).

(72) Inventors: CHEN, Chuanshi; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). GUO, Jingru;

MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). ZHOU, Zhangyan; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). CAO, Wenwen; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). XIA, Xiaobo; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). YING, Qianlan; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). WANG, Rongzhao; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US). CHEN, Gaojun; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

(74) Agent: CHATTERJEE, Aaron C. et al.; MICROSOFT TECHNOLOGY LICENSING, LLC, One Microsoft Way, Redmond, Washington 98052-6399 (US).

(54) Title: VIDEO RECORDING PROCESSING

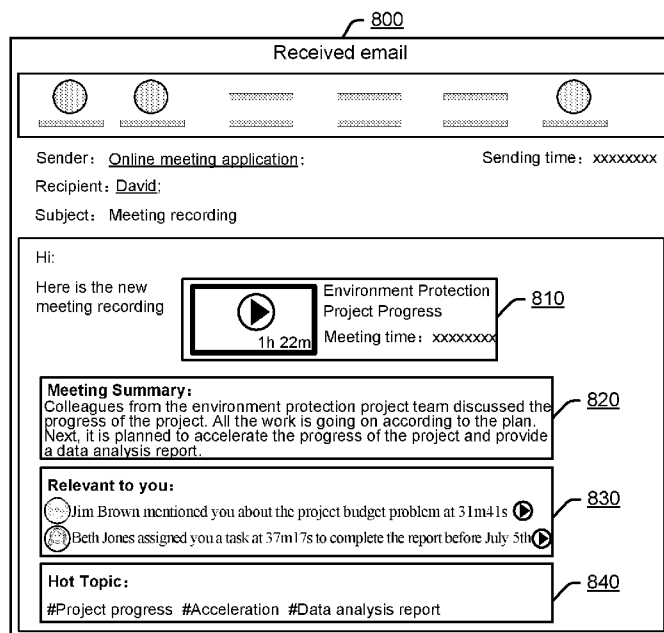


FIG 8

(57) Abstract: The present disclosure proposes methods, apparatuses, computer program products and non-transitory computer-readable mediums for processing a video recording of a target application. A video recording of the target application may be obtained. Multi-modal data of the video recording may be obtained, the multi-modal data comprising at least one of speech transcript, video, image, text and event information. A multi-modal feature of the video recording may be generated based on the multi-modal data, the multi-modal feature comprising at least one of a speech transcript feature, a video feature, an image feature, a text feature and an event feature. Target content associated with the video recording may be determined based at least on the multi-modal feature.

(81) Designated States (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*

VIDEO RECORDING PROCESSING

BACKGROUND

Video is a long, linear and experience-isolated content format. This leads to a challenge in the consuming and collaborating of a video. For example, due to the nature of a video itself, great efforts may be required to perform processing such as format conversion, editing, content extraction, etc., on the video.

SUMMARY

This Summary is provided to introduce a selection of concepts that are further described below in the Detailed Description. It is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used to limit the scope of the claimed subject matter. Embodiments of the present disclosure propose methods, apparatuses, computer program products and non-transitory computer-readable mediums for processing a video recording of a target application. A video recording of the target application may be obtained. Multi-modal data of the video recording may be obtained, the multi-modal data comprising at least one of speech transcript, video, image, text and event information. A multi-modal feature of the video recording may be generated based on the multi-modal data, the multi-modal feature comprising at least one of a speech transcript feature, a video feature, an image feature, a text feature and an event feature. Target content associated with the video recording may be determined based at least on the multi-modal feature.

It should be noted that the above one or more aspects comprise the features hereinafter fully described and particularly pointed out in the claims. The following description and the drawings set forth in detail certain illustrative features of the one or more aspects. These features are only indicative of the various ways in which the principles of various aspects may be employed, and this disclosure is intended to include all such aspects and their equivalents.

BRIEF DESCRIPTION OF THE DRAWINGS

The disclosed aspects will hereinafter be described in conjunction with the appended drawings that are provided to illustrate and not to limit the disclosed aspects.

FIG.1 illustrates an exemplary process for processing a video recording of a target application according to an embodiment.

FIG.2 illustrates an exemplary process for text summary generation according to an embodiment.

FIG.3 illustrates an exemplary process for video summary generation according to an embodiment.

FIG.4 illustrates an exemplary process for hot topic detection and hot topic video clip extraction according to an embodiment.

FIG.5 illustrates an exemplary process for mentioned moment description generation and

mentioned moment video clip extraction according to an embodiment.

FIG.6 illustrates an exemplary process for task description generation and task video clip extraction according to an embodiment.

FIG.7 illustrates an exemplary user interface of a target application.

5 FIG.8 illustrates an example of providing a prompt of target content according to an embodiment.

FIG.9 illustrates an exemplary user interface of a target application according to an embodiment.

FIG.10 illustrates an exemplary user interface of a target application according to an embodiment.

FIG.11 illustrates an example of providing a sharing message card according to an embodiment.

FIG.12 illustrates an example of updating a sharing message card according to an embodiment.

10 FIG.13 illustrates a flowchart of an exemplary method for processing a video recording of a target application according to an embodiment.

FIG.14 illustrates an exemplary apparatus for processing a video recording of a target application according to an embodiment.

15 FIG.15 illustrates an exemplary apparatus for processing a video recording of a target application according to an embodiment.

DETAILED DESCRIPTION

The present disclosure will now be discussed with reference to several example implementations.

It should be understood that these implementations are discussed only for enabling those skilled in the art to better understand and thus implement the embodiments of the present disclosure, rather than suggesting any limitations on the scope of the present disclosure.

20 Video recordings of applications are a common type of video. For example, during the running of an application, a video recording of the application may be obtained through recording a user interface of the application presented on a screen, an audio generated in the application, etc. Restricted by characteristics of a video itself, interactions with a video recording are usually limited to simply playing the video recording, artificially and restrictedly editing the video recording, etc., and it is difficult for people to quickly find desired information or interested content from the video recording.

Embodiments of the present disclosure propose to automatically perform effective processing on a video recording of a target application so as to determine various target contents. Herein, a target application may refer to various applications from which video recordings are generated. The term "application" may broadly cover software, program, client, Web application, widget, plug-in, etc. Exemplarily, a target application may include, e.g., an online meeting application, a video chatting application, a game application, a virtual reality (VR) application, a Meta-verse application, or any other application capable of generating a video recording. Moreover, herein, target content may refer to various desired information or interested content obtained or generated from a video

recording.

In some aspects, the embodiments of the present disclosure may obtain multi-modal data including data in multiple different modalities from a video recording of a target application, generate a multi-modal feature based on the multi-modal data, and utilize the multi-modal feature for determining various types of target content associated with the video recording. The determined target content may include, e.g., at least one of text summary, video summary, hot topic, hot topic video clip, mentioned moment description, mentioned moment video clip, task description, task video clip, etc.

According to the embodiments of the present disclosure, a target application may automatically determine target content associated with a video recording, so that a user can immersively consume the target content, perform target content-based collaboration, etc., in the target application. Benefiting from the determined target content, the user can easily know or navigate to key information, interested content, etc., in the video recording, and can conveniently share or collaborate on the target content with other users. Thus, the embodiments of the present disclosure can significantly improve intelligence level and user experience of the target application.

It should be understood that although multiple parts of the following discussion take an online meeting application as an example of a target application and take a meeting video recording of an online meeting application as an example of a video recording of a target application, the embodiments of the present disclosure are not limited to the scenario of the online meeting application, but may also be adopted in any other types of target application in a similar approach. FIG.1 illustrates an exemplary process 100 for processing a video recording of a target application according to an embodiment. The process 100 may be performed in a target application 102.

At 110, a video recording of the target application 102 may be obtained. In one case, the video recording may be recorded by a user of the target application 102 through operating the target application 102. In one case, the video recording may be automatically recorded by the target application 102. In any case, at 110, the target application 102 may extract a file of the video recording from a storage space storing the video recording by any approach.

At 120, multi-modal data of the video recording may be obtained. The multi-modal data may include data in multiple modalities in the video recording. The process 100 may more accurately determine target content through adopting the multi-modal data.

In an implementation, the multi-modal data may include speech transcript. A video recording may include speech from a user or a speaker. For example, a speaker in an online meeting application, a video chatting application, etc., may be a participant, a speaker in a game application, a virtual reality application, a Meta-verse application, etc., may be a game character or a player, etc. A speech transcript may refer to a file containing text converted from speech. A speech transcript

may be a set of multiple speech transcript entries in chronological order, and each speech transcript entry may include an identification of a speaker and text corresponding to speech from the speaker. An exemplary speech transcript entry is <Jim Brown: "I will introduce the progress of the project ...">, wherein "Jim Brown" is an identification of a speaker, such as the name of the speaker, and "I will introduce the progress of the project ..." is the text corresponding to the speech from Jim Brown. Moreover, each speech transcript entry may also contain time information for indicating a time point at which the speech occurs. It should be understood that the embodiments of the present disclosure are not limited to any specific form of speech transcript. The operation of obtaining multi-modal data at 120 may include: forming a speech transcript through converting multiple segments of speech in the video recording into multiple corresponding segments of text and identifying a speaker, time information, etc., of each segment of speech. It should be understood that the embodiments of the present disclosure are not limited to any specific technique for forming a speech transcript, e.g., the embodiments of the present disclosure may adopt any known speech recognition technique, speaker recognition technique, etc.

In an implementation, the multi-modal data may include video. A video recording may include at least a video which records a visual presentation of a user interface of the target application 102. Taking an online meeting application as an example, when multiple meeting participants are having an online meeting, a user interface of the online meeting application may provide a virtual workspace for the meeting participants, and accordingly, a meeting video recording may include a video which records the user interface over time.

In an implementation, the multi-modal data may include image, e.g., an image contained in a video frame of a video recording. Herein, an image may refer to various image elements presented in a user interface of a target application. Taking an online meeting application as an example, an image may include, e.g., a user avatar, an image presented in a shared screen window, an image presented in a chat window, etc. Taking a game application as an example, an image may include, e.g., a game character image, a game scene image, a player avatar, etc. The operation of obtaining multi-modal data at 120 may include: identifying and extracting an image from the video recording. It should be understood that the embodiments of the present disclosure are not limited to any specific technique for identifying and extracting an image.

In an implementation, the multi-modal data may include text, e.g., text contained in a video frame of a video recording. Herein, a text may refer to various text elements presented in a user interface of a target application. Taking an online meeting application as an example, a text may include, e.g., text presented in a shared screen window, chat text presented in a chat window, etc. The operation of obtaining multi-modal data at 120 may include: identifying and extracting text from the video recording. It should be understood that the embodiments of the present disclosure are

not limited to any specific technique for recognizing and extracting text, e.g., the embodiments of the present disclosure may adopt any known optical character recognition (OCR) technique, etc.

In an implementation, the multi-modal data may include event information. An event may refer to a usage instance occurring in a target application, and event information may refer to various

5 types of information about the event. Accordingly, a video recording may be associated with a specific event. Event information may include, e.g., event title, event introduction, event time, a list of persons involved in an event, etc. Taking an online meeting application as an example, it is assumed that multiple users are participating in a specific online meeting A, and accordingly, the online meeting A corresponds to an event, and various types of information about the online
10 meeting A corresponds to event information. For example, the event information of the online meeting A may include a meeting title, a meeting introduction, a meeting time, a list of participants, a list of invitees, etc. Accordingly, the video recording obtained at 110 may be a video recording related to the online meeting A. It should be understood that since an event occurs in a target application, the operation of obtaining multi-modal data at 120 may include: obtaining various
15 types of event information of an event corresponding to the video recording by the target application in any approach.

It should be understood that although a variety of exemplary data that multi-modal data may include is described above, the embodiments of the present disclosure are not limited to obtaining any one or more types of the data at 120, and any other type of data may also be obtained.

20 At 130, a multi-modal feature of the video recording may be generated based on the obtained multi-modal data. A multi-modal feature is a feature used for characterizing multi-modal data, which may be subsequently used by a predetermined machine learning model, neural network, etc. for determining target content. Different types of multi-modal feature may be generated for different types of multi-modal data, respectively. For example, the multi-modal feature may
25 include a speech transcript feature, a video feature, an image feature, a text feature, an event feature, etc. corresponding to speech transcript, video, image, text, event information, etc. in the multi-modal data. In an implementation, the multi-modal feature may be generated by performing an encoding process on the multi-modal data. For example, a corresponding encoder may be applied to data in each modality so as to obtain an encoded feature corresponding to the data in
30 this modality.

At 140, target content associated with the video recording may be determined based at least on the multi-modal feature. For example, multiple types of target content may be respectively determined based on combinations of different types of multi-modal feature. The determination of the target content may be performed at least through adopting a machine learning model, a
35 neural network, etc., specific to the target content.

In an implementation, the target content may include a text summary. Accordingly, the determination of the target content at 140 may include generating a text summary of the video recording based at least on the multi-modal feature. A text summary is a description of key content in a video recording, which may help a user to quickly understand the key content in the video recording in the form of text. Various types of text summary may be generated, e.g., extractive summary, abstractive summary, etc. An extractive summary is generated based at least on a speech transcript feature, which includes text converted from speech corresponding to key content in a video recording. An extractive summary aims to reflect key content in a video recording with text corresponding to actual speech from a speaker. An abstractive summary is generated based at least on a speech transcript feature or an extractive summary, wherein the abstractive summary includes generalized natural language text and the text may contain words, phrases, etc., that are not in the speech transcript. An abstractive summary aims to reflect key content in a video recording with more readable text.

In an implementation, the target content may include a video summary. Accordingly, the determination of the target content at 140 may include generating a video summary of the video recording based at least on the multi-modal feature. A video summary is a brief visual summary of a video recording, which summarizes content of the video recording through selecting the most representative, informative and important video clips from the video recording. For example, a video summary may be formed by at least a part of video frames in a video recording. A video summary is a video description of key content in a video recording, which may help a user to quickly understand the key content in the video recording in the form of video clip.

In an implementation, the target content may include a hot topic and/or a hot topic video clip. Accordingly, the determination of the target content at 140 may include detecting at least one hot topic in the video recording and/or extracting at least one hot topic video clip associated with the at least one hot topic based at least on the multi-modal feature. A hot topic is a topic that is frequently mentioned in a video recording and has high attention and importance. A hot topic may be in the form of, e.g., word, phrase, etc. A hot topic video clip associated with a hot topic is a video recording clip in which the hot topic appears in a video recording. For example, in a hot topic video clip, utterances from a specific speaker contain the hot topic. A hot topic is in the form of text, and a hot topic video clip is in the form of video recording clip, both of which may help a user to intuitively and quickly understand key content in a video recording.

In an implementation, the target content may include a mentioned moment description and/or a mentioned moment video clip. Accordingly, the determination of the target content at 140 may include: generating a mentioned moment description and/or extracting a mentioned moment video clip based at least on the multi-modal feature. In some cases, in a video recording, at a specific

time point, a speaker may mention another target user in a spoken utterance, and accordingly, this time point may correspond to a mentioned moment. A mentioned moment description is a text describing the case that a target user is mentioned. A mentioned moment description may include, e.g., an identification of a speaker, a mentioned moment, an identification of a target user, a mentioned item, etc. A mentioned moment description may be expressed through a natural language sentence. A mentioned moment video clip is a video recording clip at a mentioned moment in a video recording. A mentioned moment description is in the form of text, and a mentioned moment video clip is in the form of video recording clip, both of which may help a user to intuitively and quickly understand the case that a target user is mentioned in a video recording.

In an implementation, the target content may include a task description and/or a task video clip. Accordingly, the determination of the target content at 140 may include: generating a task description and/or extracting a task video clip based at least on the multi-modal feature. In some cases, in a video recording, at a specific time point, a speaker may mention a task associated with a target user in a spoken utterance, and accordingly, this time point may correspond to a task moment, wherein the speaker may be the same as or different from the target user. For example, if the speaker is different from the target user, the speaker may have requested or assigned the task to the target user in the spoken utterance. For example, if the speaker is the same as the target user, the speaker or user may have committed or accepted the task in the spoken utterance. A task description is a text which describes a task associated with a target user. A task description may include, e.g., an identification of a speaker, a task moment, an identification of a target user, task content, etc. A task description may be expressed through a natural language sentence. A task video clip is a video recording clip at a task moment in a video recording. A task description is in the form of text, and a task video clip is in the form of video recording clip, both of which may help a user to intuitively and quickly understand a task associated with a target user in a video recording.

It should be understood that although a variety of exemplary target content is described above, the embodiments of the present disclosure are not limited to determining any one or more types of the target content at 140, and any other type of target content may also be determined.

At 150, a prompt of the target content may be provided and/or the target content may be presented. In an aspect, at 150, a prompt regarding the target content determined at 140 may be provided to a user of the target application 102 in various approaches. The user who will receive the prompt may be a user related to the video recording or a person involved in an event, e.g., a participant, an invitee, etc., involved in an event associated with the video recording. In an approach, the target application may invoke an email application to generate a prompt email and send the prompt email

to the user. The prompt email may include an event introduction associated with the video recording and the target content associated with the video recording. Thus, the user who has received the prompt email may easily learn key content in the video recording through the prompt email without the need of watching the entire video recording. Preferably, the prompt email may be specific to the target user, and the target user is a user mentioned in the video recording, a user associated with a task, etc. Thus, the target user who is a recipient of the prompt email may conveniently know the case that the target user is mentioned in the video recording, a task associated with the target user, etc., through the prompt email. In another approach, the target application may invoke chatting tool software to generate a prompt message and send the prompt message to a user. The prompt message may include content similar to the prompt email as described above. In another approach, instead of invoking an email application, chatting tool software, etc., the target application 102 may set up a video recording hub which may store multiple video recordings respectively associated with different events and target content determined for each video recording. After determining target content associated with a specific video recording at 140, the video recording and the associated target content may be saved to the video recording hub, and a prompt notification may be sent to a user in the target application 102 so as to inform the user that the user may go to the video recording hub to view the video recording and the target content. It should be understood that the embodiments of the present disclosure are not limited to providing a prompt about target content in any one or more of the exemplary approaches as described above, but may also provide a prompt about target content in any other approach. Moreover, it should be understood that the e-mail application, the chatting tool software, etc. as described above may be in a unified integrated software environment with the target application, e.g., these applications and software are different functions provided by the integrated software environment, or the e-mail application, the chatting tool software, etc. may be independent of the target application, e.g., they may be accessed by the target application as third-party applications.

In an aspect, at 150, the target content may be presented to a user of the target application 102. The user may be a user related to the video recording or a person involved in an event. In an implementation, the target application 102 may include a specific target content presentation user interface. In the target content presentation user interface, a user may intuitively and conveniently access a video recording, target content, etc. The target content presentation user interface may be displayed in the target application in response to a predetermined operation by a user on a prompt of target content. For example, when a user clicks on a specific region in a prompt email, prompt message, prompt notification, etc., the displaying of the target content presentation user interface in the target application may be triggered. It should be understood that the embodiments of the

present disclosure are not limited to any specific design, layout, etc. of the target content presentation user interface.

Assuming that a user wants to share specific target content to at least one recipient in the target content presentation user interface, the process 100 may optionally include: at 160, in response to receiving a request of sharing the target content to at least one recipient, generating a sharing message card associated with the target content; and at 170, providing the sharing message card to the at least one recipient. Exemplarily, the target application may invoke chatting tool software to provide the sharing message card to the recipient. The sharing message card may be an information card especially designed for sharing target content, which may include, e.g., an identification of a user who is a sharing initiator, a comment from the sharing initiator, description and link of the shared target content, etc. Moreover, optionally, the sharing message card may also have a comment function. For example, the sharing initiator, recipients, etc., of the sharing message card may post comments in the sharing message card, and the sharing message card has a specific comment region for dynamically collecting comments. Thus, the sharing initiator and the recipients may conveniently conduct discussions, etc. on the shared target content in the sharing message card.

It should be understood that all the operations or steps in the process 100 as described above in connection with FIG.1 are exemplary, and depending on specific application scenarios and requirements, the process 100 may include more or less operations or steps, and the embodiments of the present disclosure will cover changes to the process 100 in any approach. Moreover, herein, the word "user" may also be used interchangeably with user of a target application, person involved in an event, speaker, participant, invitee, player, etc.

FIG.2 illustrates an exemplary process 200 for text summary generation according to an embodiment. The process 200 is an exemplary implementation of the step 120 to step 140 in FIG.1.

It is assumed that a speech transcript 202 in multi-modal data of a video recording has been obtained through the step 120 in FIG.1. According to the process 200, the speech transcript 202 may be further divided into multiple transcript segments 204. In an implementation, transcript segment division may be performed sentence by sentence, so that each transcript segment includes a sentence converted from speech. For example, each transcript segment may include an identification of a speaker and a sentence text corresponding to speech from the speaker. Accordingly, transcript segments may be a finer-grained division of a speech transcript entry, e.g., a speech transcript entry may be divided into multiple transcript segments on the basis of sentence. Through dividing a speech transcript into multiple transcript segments and performing subsequent processing transcript segment-by-transcript segment, a text summary may be generated more accurately. Moreover, each transcript segment may also contain time information for indicating a

time point at which a speaker said a sentence in the transcript segment. The embodiments of the present disclosure may adopt any known technique, e.g., audio pause detection, shot boundary detection, etc., for performing transcript segment division. It should be understood that the embodiments of the present disclosure are not limited to adopting any specific technique for performing transcript segment division.

The process 200 may adopt an estimator 210 for analyzing each transcript segment, so as to identify whether the transcript segment should be included in an extractive summary. The estimator 210 may be a previously-trained machine learning model, neural network, etc., for example, it may be a transformer-based natural language processing model. Exemplarily, the estimator 210 may include an encoder 212 and a decoder 214. The encoder 212 may encode each input transcript segment so as to obtain a corresponding transcript segment feature. The process of utilizing the encoder 212 for generating a transcript segment feature based on a transcript segment may be considered as an exemplary implementation of generating a speech transcript feature at the step 130 in FIG.1. For example, multiple transcript segment features respectively corresponding to the multiple transcript segments 204 together form a speech transcript feature. The decoder 214 may determine whether the current transcript segment should be included in the extractive summary based on a speech transcript feature. The process of utilizing the decoder 214 for determining a transcript segment to be included in an extractive summary based on a transcript segment feature may be considered as an exemplary implementation of generating a text summary at the step 140 in FIG.1. The estimator 210 may be trained for identifying, from the multiple transcript segments 204, a transcript segment subset that can reflect key content in the video recording and thus should be included in an extractive summary. Accordingly, the estimator 210 will output an extractive summary 220 formed from the identified transcript segment subset.

An exemplary extractive summary may be: [..., <Jim Brown: "By the end of June, we have completed the first phase of work of this project">, <Beth Jones: "In the next step, it is needed to accelerate the completion of a project report">, ...]. The exemplary extractive summary includes at least a text corresponding to speech from several speakers that can reflect key content in the video recording, e.g., the text "By the end of June, we have completed the first phase of work of this project" corresponding to an utterance spoken by Jim Brown, the text "In the next step, it is needed to accelerate the completion of a project report" corresponding to an utterance spoken by Beth Jones, etc.

According to the process 200, optionally, an abstractive summary 240 may be generated based on the extractive summary 220. In an implementation, a generator 230 may be adopted for generating the abstractive summary 240 based on the extractive summary 220. The generator 230 may be a previously-trained machine learning model, neural network, etc., for example, it may be a

sequence-to-sequence model. Exemplarily, the generator 230 may include an encoder 232 and a decoder 234. The encoder 232 may encode each transcript segment in the abstractive summary 240 so as to obtain a corresponding transcript segment feature. The decoder 234 may generate a corresponding natural language sentence based on a transcript segment feature and include the natural language sentence into the abstractive summary. Preferably, each transcript segment in the extractive summary 220 has a corresponding natural language sentence in the abstractive summary 240. The process of utilizing the generator 230 for generating the abstractive summary 240 based on the extractive summary 220 may be considered as a further exemplary implementation of generating a text summary at the step 140 in FIG.1. As an example, assuming that an extractive summary includes at least a transcript segment <Jim Brown: "By the end of June, we have completed the first phase of work of this project">, an abstractive summary may include at least a natural language sentence corresponding to the transcript segment, e.g., <Jim Brown talked that the first phase of work has been completed by the end of June>.

Alternatively, the process 200 may also directly generate the abstractive summary 240 for the multiple transcript segments 204. In this case, the generator 230 may be trained for generating multiple natural language sentences in an abstractive summary directly based on the multiple transcript segments 204. Accordingly, the encoder 232 can encode each input transcript segment into a corresponding transcript segment feature, so as to obtain the entire speech transcript feature. The process of utilizing the encoder 232 for generating a speech transcript feature based on multiple transcript segments may be considered as an exemplary implementation of generating a speech transcript feature at the step 130 in FIG.1. The decoder 234 may generate multiple natural language sentences in an abstractive summary based on a speech transcript feature. The process of utilizing the decoder 234 for generating the abstractive summary 240 based on a speech transcript feature may be considered as an exemplary implementation of generating a text summary at the step 140 in FIG.1.

It should be understood that either one or both of the extractive summary 220 and the abstractive summary 240 may be taken as a text summary of the video recording. Moreover, the embodiments of the present disclosure are not limited to generating a text summary through the estimator 210 and the generator 230 as described above, and are not limited to any specific technique for training the estimator 210 and the generator 230.

According to the process 200, optionally, the extractive summary 220 and/or the abstractive summary 240 may be calibrated at 250 with at least a text 206 in the multi-modal data, e.g., correcting or replacing incorrect or misspelled words, etc. in the extractive summary 220 and/or the abstractive summary 240 with words in the text 206. It is assumed that the text 206 in the multi-modal data of the video recording has been obtained through the step 120 in FIG.1. For

example, the text 206 may be a text presented in a shared screen window, a chat text presented in a chat window, etc. Taking an online meeting application as an example, a user interface may include a shared screen window in which meeting participants may share a specific document or screen, and accordingly, the text 206 may be presented in the shared screen window. Moreover, for example, the user interface of the online meeting application may also include a chat window in which meeting participants may chat through inputting text, images, etc., and accordingly, the text 206 may be presented in the chat window. The text 206 may contain words associated with speech from a speaker, and these words may be words that the speaker really intended to express. Thus, if the extractive summary 220 and/or the abstractive summary 240 contain incorrect or misspelled words, the incorrect or misspelled words may be modified or replaced by words in the text 206. Preferably, proper words such as technical terms, professional terms, personal names, etc. may be identified from the text 206 through techniques such as the named entity recognition (NER). Then, the proper words identified from the text 206 may be used for calibrating corresponding words in the extractive summary 220 and/or the abstractive summary 240. Through performing calibration at 250, expressions in the extractive summary 220 and/or the abstractive summary 240 may be more accurate.

It should be understood that all the operations or steps in the process 200 as described above in connection with FIG.2 are exemplary, and depending on specific application scenarios and requirements, the process 200 may include more or less operations or steps, and the embodiments of the present disclosure will cover changes to the process 200 in any approach.

FIG.3 illustrates an exemplary process 300 for video summary generation according to an embodiment. The process 300 is an exemplary implementation of the step 120 to step 140 in FIG.1. It is assumed that at least one of a speech transcript 302, a video 304, an image 306 and a text 308 in multi-modal data of a video recording has been obtained through the step 120 in FIG.1.

An encoder 310 may encode the speech transcript 302 so as to obtain a speech transcript feature. Preferably, the encoder 310 may encode each transcript segment in the speech transcript 302 in an approach similar to the encoder 212 in FIG.2 to obtain a corresponding transcript segment feature, and utilize multiple transcript segment features for forming a speech transcript feature. Preferably, the encoder 310 may encode contextual information in textual modality, which may be a sentence-level encoder, a hierarchical document-level encoder, etc. Exemplarily, the encoder 310 may be based on recurrent neural network (RNN) architecture, which may adopt, e.g., a long short-term memory (LSTM) unit, a gated recurrent unit (GRU), etc.

An encoder 320 may encode the video 304 to obtain a video feature. In an implementation, the encoder 320 may encode each video frame in the video 304 to obtain a corresponding video frame feature, and utilize multiple video frame features for forming a video feature. Moreover, the

encoding of the video 304 by the encoder 320 may also include a sequential encoding process. Preferably, the encoder 320 may be based on recurrent neural network (CNN) architecture, which may capture contextual information of a video frame.

An encoder 330 may encode the image 306 to obtain an image feature. Preferably, the encoder 330 may be based on CNN architecture, which is similar to the encoder 320.

An encoder 340 may encode the text 308 to obtain a text feature. Preferably, the encoder 340 may encode contextual information in textual modality. Exemplarily, the encoder 340 may be based on RNN architecture, which is similar to the encoder 310.

The process of utilizing the encoders 310, 320, 330 and 340 for respectively generating the speech transcript feature, the video feature, the image feature and the text feature may be considered as an exemplary implementation of generating a multi-modal feature at the step 130 in FIG.1.

At 350, a merged feature may be obtained through merging at least one of the speech transcript feature, the video feature, the image feature and the text feature. Various merging strategies may be adopted for performing the merging at 350. In a merging strategy which is based on feature concatenation, vector representations of a speech transcript feature, a video feature, an image feature and a text feature may be directly concatenated so as to obtain a merged feature. In a merging strategy which is based on weighted summation, vector representations of a speech transcript feature, a video feature, an image feature and a text feature may be imposed by corresponding weights and summed, so as to obtain a merged feature. In an attention-based merging strategy, an attention mechanism may be utilized for combining a speech transcript feature, a video feature, an image feature and a text feature so as to enhance influences from some of these features and weaken influences from other features, thereby eliminating noises and focusing on relevant information. It should be understood that the embodiments of the present disclosure are not limited to any one or more of the merging strategies as discussed above, and may also adopt any other merging strategies.

According to the process 300, at least a part of video frames may be further selected from the video recording based on the merged feature so as to form a video summary 362. For example, a video summary generator 360 may be adopted for generating the video summary 362 based on the merged feature. The video summary generator 360 may be a machine learning model, neural network, etc. trained for selecting representative, informative and important video frames from a video recording so as to form a video summary. The video summary generator 360 may be implemented based on various techniques. In an implementation, the video summary generator 360 may be implemented based on a sequence generation network. For example, the video summary generator 360 may be a transformer model. In an implementation, the video summary generator 360 may be implemented based on a graph network. The process of generating a video

summary may be generalized as a graph analysis problem. The graph-based technique may effectively address the shortcomings of traditional sequence models in long-distance dependency capturing. Moreover, a graph network may also better preserve video content and short-level dependencies during a summary generation process. In an implementation, the video summary generator 360 may be implemented based on a reinforcement learning network. The process of generating a video summary may be generalized as a sequential decision-making process. The reinforcement learning network may predict a probability for each video frame, which indicates how likely this video frame will be selected. Then, the reinforcement learning network may take actions for selecting multiple video frames based on probability distribution, so as to form a video summary. It should be understood that the embodiments of the present disclosure are not limited to any one or more implementations of the video summary generator as described above, and may also adopt any other approach for implementing the video summary generator.

The merging at 350, the video summary generation at 360, etc., may be considered as an exemplary implementation of determining a target content at the step 140 in FIG.1.

It should be understood that all the operations or steps in the process 300 as described above in connection with FIG.3 are exemplary, and depending on specific application scenarios and requirements, the process 300 may include more or less operations or steps, and the embodiments of the present disclosure will cover changes to the process 300 in any approach. For example, although the process 300 involves generating a video summary with the speech transcript 302, the video 304, the image 306 and the text 308, the video summary may also be generated with only one or more of the speech transcript 302, the video 304, the image 306 and the text 308. Moreover, preferably, the process 300 may also include temporally aligning the speech transcript 302, the video 304, the image 306 and the text 308, so that multi-modal data at a time point corresponding to a video frame may be considered comprehensively when determining whether the video frame should be selected to add into a video summary. Moreover, although various encoders and video summary generators are described above separately, these encoders and video summary generators may be jointly trained, and the embodiments of the present disclosure are not limited to any specific training approach.

FIG.4 illustrates an exemplary process 400 for hot topic detection and hot topic video clip extraction according to an embodiment. The process 400 is an exemplary implementation of the step 120 to step 140 in FIG.1.

It is assumed that at least one of a speech transcript 402, a video 404 and an event information 406 in multi-modal data of a video recording has been obtained through the step 120 in FIG.1.

At 410, candidate topic identification may be performed on the speech transcript 402, so as to identify multiple candidate topics 412 from the speech transcript 402. In an implementation,

candidate topics may be identified based on predetermined rules. For example, based on occurrence frequencies of words or phrases, multiple words or phrases with the highest occurrence frequencies may be selected from the speech transcript 402 as candidate topics. The embodiments of the present disclosure are not limited to the approaches of identifying candidate topics as described above, but may also adopt any other approach for identifying candidate topics from the speech transcript 402.

An encoder 420 may encode the speech transcript 402 to obtain a speech transcript feature. The implementation of the encoder 420 may be similar to the encoder 310 in FIG.3.

An encoder 430 may encode the video 404 to obtain a video feature. The implementation of the encoder 430 may be similar to the encoder 320 in FIG.3.

An encoder 440 may encode the event information 406 to obtain an event feature, e.g., a vector representation of the event information. Preferably, the encoder 440 may be based on RNN architecture. Moreover, preferably, the event information 406 encoded by the encoder 440 may be unstructured data including event title, event introduction, etc., such as, meeting title, meeting introduction, etc. in the scenario of online meeting application.

The process of utilizing the encoders 420, 430 and 440 for respectively generating a speech transcript feature, a video feature and an event feature may be considered as an exemplary implementation of generating a multi-modal feature at the step 130 in FIG.1.

At 450, hot topic selection may be performed so as to select at least one hot topic 454 from the candidate topics 412 based on at least one of a speech transcript feature, a video feature and an event feature. For example, the hot topic selection at 450 may adopt at least a scoring model 452. The scoring model 452 may be trained for taking the candidate topics 412 and at least one of a speech transcript feature, a video feature and an event feature as inputs, and outputting a score for each candidate topic. The scoring model 452 may be, e.g., a deep neural network-based model.

After obtaining a score for each candidate topic, at least one candidate topic with the highest score may be selected as the hot topic 454.

Preferably, the process 400 may further include performing video clip extraction at 460, so as to extract at least one hot topic video clip 462 associated with the at least one hot topic 454 from the video recording. In an implementation, for a hot topic, a transcript segment containing the hot topic and a start time point of the transcript segment may be identified. Then, a video clip within a time range at least containing the start time point may be extracted from the video recording as a hot topic video clip associated with the hot topic.

The hot topic selection at 450, the video clip extraction at 460, etc. may be considered as an exemplary implementations of determining target content at the step 140 in FIG.1.

It should be understood that all the operations or steps in the process 400 described above in

conjunction with FIG.4 are exemplary, and depending on specific application scenarios and requirements, the process 400 may include more or less operations or steps, and the embodiments of the present disclosure will cover changes to the process 400 in any approach. For example, although the process 400 involves utilizing the speech transcript 402, the video 404 and the event information 406 for detecting a hot topic, the hot topic may be detected only with one or more of the speech transcript 402, the video 404 and the event information 406. Moreover, preferably, the process 400 may also include temporally aligning the speech transcript 402 and the video 404. Moreover, although various encoders and scoring models are described above separately, these encoders and scoring models may be jointly trained, and the embodiments of the present disclosure are not limited to any specific training approach.

FIG.5 illustrates an exemplary process 500 for mentioned moment description generation and mentioned moment video clip extraction according to an embodiment. The process 500 is an exemplary implementation of the step 120 to step 140 in FIG.1.

It is assumed that at least one of a speech transcript 502 and event information 504 in multi-modal data of a video recording has been obtained through the step 120 in FIG.1.

An encoder 510 may encode the speech transcript 502 to obtain a speech transcript feature. The implementation of the encoder 510 may be similar to the encoder 310 in FIG.3.

An encoder 520 may encode the event information 504 to obtain an event feature. The implementation of the encoder 520 may be similar to the encoder 440 in FIG.4. Preferably, the event information 504 encoded by the encoder 520 may be unstructured data including event title, event introduction, etc.

The process of utilizing the encoders 510 and 520 for respectively generating a speech transcript feature and an event feature may be considered as an exemplary implementation of generating a multi-modal feature at the step 130 in FIG.1.

At 530, detection of transcript segment mentioning a target user may be performed so as to detect at least one transcript segment 534 mentioning the target user in the speech transcript 502 based on at least one of a speech transcript feature and an event feature. For example, the transcript segment detection at 530 may adopt at least a token labeling model 532. The token labeling model 532 may be trained for detecting a transcript segment containing a person name in a text corresponding to a speech based on at least one of a speech transcript feature and an event feature. As an example, for a transcript segment <Jim Brown: "I would like to thank David and his team for doing a great job on solving the project budget problem">, the token labeling model 532 may detect the person names "Jim Brown" and "David" from the transcript segment, wherein "Jim Brown" is a speaker and "David" is a target user mentioned by "Jim Brown". It should be understood that the embodiments of the present disclosure are not limited to implementing and

training the token labeling model 532 through adopting any specific approach.

At 540, description generation may be performed so as to generate at least one mentioned moment description 542 corresponding to the at least one transcript segment 534 based on the at least one transcript segment 534 and the event information 504. Preferably, the event information 504 used for generating a mentioned moment description may be structured data including, e.g., event time, a list of person involved in an event, etc., such as, meeting time, participant list, invitee list, etc. in the scenario of an online meeting application. In an implementation, firstly, a person name detected in the transcript segment 534 may be compared with the list of persons involved in an event in the event information 504 so as to determine the complete name of the person. Typically, the list of persons involved in an event may include full names of persons involved in the current event. For example, assuming that a person name "David" is detected from the transcript segment, and a person name "David Wilson" is included in the list of persons involved in an event, it may be determined that the full name of the mentioned target user "David" is "David Wilson". Similarly, a full name of a speaker may be determined. A mentioned moment corresponding to the transcript segment 534 may be determined. Then, a mentioned moment description may be generated with at least the transcript segment 534, the full names of the target user and the speaker, the mentioned moment, etc. For example, a previously-trained sentence generation model may be adopted for generating a mentioned moment description expressed in a natural language sentence. The mentioned moment description may include, e.g., an identification of a speaker, a mentioned moment, an identification of a target user, a mentioned item, etc. Assuming that a transcript segment mentioning a target user is <Jim Brown: "I would like to thank David and his team for doing a great job on solving the project budget problem">, the full name of the target user is "David Wilson", and a mentioned moment is the 31st minute and 41st second (i.e., 31m41s) of a video recording, a generated mentioned moment description may be "Jim Brown mentioned David Wilson about the project budget problem at 31m41s", wherein "project budget problem" is a mentioned item. Optionally, if the mentioned moment description is to be provided to the target user David himself, the generated mentioned moment description may also be "Jim Brown mentioned you about the project budget problem at 31m41s".

According to the process 500, optionally, video clip extraction may be performed at 550 so as to extract at least one mentioned moment video segment 552 from the video recording based on the at least one transcript segment 534. In an implementation, for a transcript segment, a mentioned moment corresponding to the transcript segment may be identified. Then, a video clip within a time range at least containing the mentioned moment may be extracted from the video recording as a mentioned moment video clip associated with the transcript segment.

The detection of transcript segment mentioning a target user at 530, the description generation at

540, the video clip extraction at 550, etc. may be considered as exemplary implementations of determining target content at the step 140 in FIG.1.

It should be understood that all the operations or steps in the process 500 as described above in connection with FIG.5 are exemplary, and depending on specific application scenarios and requirements, the process 500 may include more or less operations or steps, and the embodiments of the present disclosure will cover changes to the process 500 in any approach. For example, although the process 500 involves utilizing the speech transcript 502 and the event information 504 for detecting a transcript segment mentioning a target user, a transcript segment mentioning a target user may also be detected only with one of the speech transcript 502 and the event information 504.

FIG.6 illustrates an exemplary process 600 for task description generation and task video clip extraction according to an embodiment. The process 600 is an exemplary implementation of the step 120 to step 140 in FIG.1.

It is assumed that at least one of a speech transcript 602 and an event information 604 in multi-modal data of a video recording has been obtained through the step 120 in FIG.1.

An encoder 610 may encode the speech transcript 602 to obtain a speech transcript feature. The implementation of the encoder 610 may be similar to the encoder 310 in FIG.3.

An encoder 620 may encode the event information 604 to obtain an event feature. The implementation of the encoder 620 may be similar to the encoder 440 in FIG.4. Preferably, the event information 604 encoded by the encoder 620 may be unstructured data including event title, event introduction, etc.

The process of utilizing the encoders 610 and 620 for respectively generating a speech transcript feature and an event feature may be considered as an exemplary implementation of generating a multi-modal feature at the step 130 in FIG.1.

At 630, detection of transcript segment containing a task associated with a target user may be performed so as to detect at least one transcript segment 634 containing a task associated with a target user in the speech transcript 602 based on at least one of a speech transcript feature and an event feature. For example, the transcript segment detection at 630 may adopt at least a classification model 632. The classification model 632 may be trained for detecting a transcript segment containing a task associated with a target user based on at least one of a speech transcript feature and an event feature. Exemplarily, the classification model 632 may classify an input transcript segment into one of No Task, Request Task, Commit Task, etc. A transcript segment with the No Task type does not contain any task. A transcript segment with the Request Task type may indicate that a speaker in the transcript segment requests a target user to perform a specific task, e.g., in a transcript segment <Beth Jones: "David needs to complete the report by next

Monday">, a speaker "Beth Jones" requests or assigns David a task of completing the report by next Monday. A transcript segment with the Commit Task type may indicate that a speaker in this transcript segment is a target user and that the speaker has committed to complete a specific task, e.g., in a transcript segment <David Wilson: "I will try to complete the report by next Monday">, a speaker "David Wilson" commits or accepts a task of completing the report by next Monday. It should be understood that the embodiments of the present disclosure are not limited to implementing and training the classification model 632 through adopting any specific approach, and are not limited to classifying a transcript segment into the exemplary types as described above. At 640, description generation may be performed so as to generate at least one task description 642 corresponding to the at least one transcript segment 634 based on the at least one transcript segment 634 and the event information 604. Preferably, the event information 604 used for generating a task description may be structured data including, e.g., event time, a list of persons involved in an event, etc. In an implementation, similar to the step 540 in FIG.5, firstly, a person name detected in the transcript segment 634 may be compared with a list of persons involved in an event in the event information 604 so as to determine the complete name of the person. A task moment corresponding to the transcript segment 634 may be determined. Then, a task description may be generated with at least the transcript segment 634, the full names of the target user and the speaker, the task moment, etc. For example, a previously-trained sentence generation model may be adopted for generating a task description expressed in a natural language sentence. The task description may include, e.g., an identification of a speaker, a task moment, an identification of a target user, a task content, etc. Assuming that a transcript segment containing a task associated with a target user is <Beth Jones: "David needs to complete the report before next Monday">, the full name of the target user is "David Wilson", and a task moment is the 37th minute 17th second (i.e., 37m17s) of a video recording, a generated task description may be "Beth Jones assigns David Wilson a task at 37m17s to complete the report before July 5th", wherein "completing the report before July 5th" is task content. Optionally, if the task description is to be provided to the target user David himself, the generated task description may also be "Beth Jones assigns you a task at 37m17s to complete the report before July 5th", wherein the name of the target user is replaced by the second person pronoun "you". It should be understood that the time "before July 5th" for completing the task included in the task content may be deduced based on the time-related expression "next Monday" in the transcript segment and the event time in the event information. For example, according to the event time on which the current event occurs, "next Monday" is deduced as indicating "July 5th".

According to the process 600, optionally, video clip extraction may be performed at 650 so as to extract at least one task video segment 652 from the video recording based on the at least one

transcript segment 634. In an implementation, for a transcript segment, a task moment corresponding to the transcript segment may be identified. Then, a video clip within a time range at least containing the task moment may be extracted from the video recording as a task video clip associated with the transcript segment.

- 5 The detection of transcript segment containing a task associated with a target user at 630, the description generation at 640, the video clip extraction at 650, etc. may be considered as exemplary implementations of determining target content at the step 140 in FIG.1.

It should be understood that all the operations or steps in the process 600 as described above in connection with FIG.6 are exemplary, and depending on specific application scenarios and requirements, the process 600 may include more or less operations or steps, and the embodiments
10 of the present disclosure will cover changes to the process 600 in any approach. For example, although the process 600 involves utilizing the speech transcript 602 and the event information 604 for detecting a transcript segment containing a task associated with a target user, a transcript segment containing a task associated with a target user may also be detected only with one of the
15 speech transcript 602 and the event information 604.

FIG.7 illustrates an exemplary user interface 700 of a target application. As an example, the target application in FIG.7 may be an online meeting application, and accordingly, the user interface 700 may be a user interface of the online meeting application presented on a screen of a terminal device of a specific user when multiple users or meeting participants are conducting an online meeting.

- 20 The user interface 700 may include a meeting title "Environment Protection Project Progress" of the current meeting shown in the top region.

The user interface 700 may include a participant region 710. A list of users participating in the meeting is shown in the participant region 710, wherein each user has a corresponding avatar or icon.

- 25 The user interface 700 may include a shared screen window 720. It is assumed that a user Jim is presenting a slide in the shared screen window 720. As shown, the slide may include image, text, etc.

The user interface 700 may include a chat window 730. The users participating in the meeting may chat in the chat window 730. A historical chat record is shown in the chat window 730, which
30 may include text, image, etc.

The users participating in the meeting may communicate through speech by turning on their respective microphones, or further communicate through video by turning on their respective cameras.

- 35 As the meeting is going on, the user interface 700, speeches from the users participating in the meeting, etc. may be recorded so as to form a video recording.

It should be understood that all the elements in the user interface 700 as described above in connection with FIG.7 are exemplary. The embodiments of the present disclosure are not limited by any of the details presented in FIG.7, and the user interface 700 may include more or less elements, may adopt any other layout approach, etc.

5 FIG.8 illustrates an example of providing a prompt of target content according to an embodiment. In FIG.8, the prompt of the target content is provided in the form of prompt email. The example of FIG.8 is a continuation of the example of FIG.7.

It is assumed that after a video recording is obtained according to the scenario in FIG.7, target content associated with the video recording is generated according to the embodiments of the present disclosure. Furthermore, the online meeting application may invoke an email application to generate a prompt email 800, and send the prompt email 800 to a user David. The user David may view the prompt email 800 in an inbox of the email application. The user David may be a participant or an invitee of the meeting.

10 The prompt email 800 may present an introduction of the meeting in a region 810, e.g., meeting title, meeting time, link of meeting video recording, etc.

The prompt email 800 may present a text summary of the video recording of the meeting in a region 820, e.g., "Colleagues from the environment protection project team discussed ...".

The prompt email 800 may present, in a region 830, content associated with the user David in the video recording, e.g., a mentioned moment description "Jim Brown mentioned you about the project budget problem at 31m41s", a task description "Beth Jones assigned you a task at 37m17s to complete the report before July 5th", etc. As shown, link icons of corresponding mentioned moment video clip and task video clip are further added behind the mentioned moment description and the task description.

20 The prompt email 800 may present hot topics in the video recording in a region 840, e.g., "Project progress", "Acceleration", "Data analysis report", etc.

It should be understood that all the elements in the prompt email 800 as described above in connection with FIG.8 are exemplary. The embodiments of the present disclosure are not limited by any of the details presented in FIG.8, and the prompt email 800 may include more or less elements, may adopt any other layout approach, etc. For example, the prompt email 800 may include more or less of the target content, may present the target content in any other approach, etc. Moreover, it should be understood that the embodiments of the present disclosure may send prompt emails to any one or more of participants or invitees of the meeting based on predetermined policies.

30 FIG.9 illustrates an exemplary user interface 900 of a target application according to an embodiment. The example of FIG.9 is a continuation of the example of FIG.8. Assuming that the

35

user David requests to further view the target content through, e.g., clicking on the presented target content, video clip link, video recording link, etc. in the prompt email 800 shown in FIG.8, a user interface 900 of the online meeting application may be displayed on a terminal device of the user David. The user interface 900 may be a target content presentation user interface designed for
5 accessing video recordings and target content.

The user interface 900 may include a play region 910. In the play region 910, video content selected by a user may be played.

The user interface 900 may include a video summary region 920 which presents video summary links.

10 The user interface 900 may include an important clip region including, e.g., a mentioned moment region 930, a task region 940, a hot topic region 950, etc. For example, a mentioned moment description and a mentioned moment video clip link are presented in the mentioned moment region 930, a task description and a task video clip link are presented in the task region 940, a hot topic and a hot topic video clip link are presented in the hot topic region 950, etc. It should be
15 understood that, in order to adapt to the limitation of display size, information about the mentioned moment description, task description, hot topic, etc. presented in the user interface 900 may be abbreviated or transformed versions obtained on the basis of the original mentioned moment description, task description, hot topic, etc. according to a predetermined strategy.

In the user interface 900, in response to a user's clicking or selecting of a video summary link, a
20 mentioned moment video clip link, a task video clip link, etc., the selected video summary or video clip may be played in the play region 910.

The user interface 900 may include a share button 960. Assuming that the user wants to share specific target content in the user interface 900 to other users or recipients, the user may select the target content to be shared and click the share button 960 to trigger a sharing process.

25 It should be understood that all the elements in the user interface 900 as described above in connection with FIG.9 are exemplary. The embodiments of the present disclosure are not limited by any of the details presented in FIG.9, and the user interface 900 may include more or less elements, may adopt any other layout approach, etc. For example, the user interface 900 may also present a text summary of the video recording, may present the target content in any other
30 approach, etc.

FIG.10 illustrates an exemplary user interface 1000 of a target application according to an embodiment. The example of FIG.10 is a continuation of the example of FIG.9. The user interface 1000 may correspond to the user interface 900 in FIG.9. Assuming that a user selects the hot topic region 950 in FIG.9 and clicks the share button 960, a sharing setting page 1010 may be then
35 presented in the user interface 1000.

The sharing setting page 1010 may be designed for enabling a user to make settings for a sharing operation. The sharing setting page 1010 may include a sharing initiator comment input region 1020 for a sharing initiator to input a comment. As shown, exemplarily, the user David, as a sharing initiator, inputs "Beth said in the meeting that the project is going well" in the sharing initiator comment input region 1020. The sharing setting page 1010 may include a shared content region 1030 that presents information about the shared target content. The sharing setting page 1010 may include a recipient specifying region 1040. A user may input or select a recipient in the recipient specifying region 1040. It is assumed that the user David has selected "Team A" as a recipient among recipient candidates in the recipient specifying region 1040. When the user clicks a "Send" button in the sharing setting page, a request to share the selected target content to the specified recipient will be generated. The embodiments of the present disclosure may further generate a sharing message card in response to the request and provide the sharing message card to the recipient on behalf of the user David.

It should be understood that all the elements in the user interface 1000 and the sharing setting page 1010 as described above in connection with FIG.10 are exemplary. The embodiments of the present disclosure are not limited by any of the details presented in FIG.10, and the user interface 1000 and the sharing setting page 1010 may include more or less elements, may adopt any other layout approach, etc. Moreover, the embodiments of the present disclosure also support a user to designate and share a specific video clip in a video recording. For example, a user may specify a start time point and an end time point of a specific video clip, and the embodiments of the present disclosure may share the specific video clip in an approach similar to the sharing of target content as described above.

FIG.11 illustrates an example of providing a sharing message card according to an embodiment. The example of FIG.11 is a continuation of the example of FIG.10.

It is assumed that a target application invokes chatting tool software to provide a sharing message card to a recipient "Team A". FIG.11 illustrates a user interface 1100 of a group chat of the recipient "Team A" in the chatting tool software. The user interface 1100 presents a sharing message card 1110 from a user David Wilson. The sharing message card 1110 is generated in response to the user's request in FIG.10 according to the embodiments of the present disclosure. The sharing message card 1110 may include information about the shared target content, e.g., a link to a hot topic video clip, a description about a hot topic "Beth Jones talked about #Project progress" etc. The sharing message card 1110 may include a comment from the sharing originator, "Beth said in the meeting that the project is going well". The sharing message card 1110 may also include a "Reply" button, so that a user receiving the sharing message card 1110 may post a comment. In this example, all the team members of the recipient "Team A" may view the sharing

message card 1110 in the group chat, and thus may utilize the reply button for posting comments. It should be understood that all the elements in the user interface 1100 and the sharing message card 1110 as described above in connection with FIG.11 are exemplary. The embodiments of the present disclosure are not limited by any of the details presented in FIG.11, and the user interface 1100 and the sharing message card 1110 may include more or less elements, may adopt any other layout approach, etc. Moreover, the embodiments of the present disclosure are not limited to any specific technique for generating a sharing message card.

FIG.12 illustrates an example of updating a sharing message card according to an embodiment. The example of FIG.12 is a continuation of the example of FIG.11. The sharing message card 1200 in FIG.12 may correspond to the sharing message card 1110 in FIG.11. According to the embodiments of the present disclosure, a sharing message card may have a comment function. As shown in FIG.12, the sharing message card 1200 includes a comment region 1210 which dynamically collects and presents comments having been received. The comment region 1210 may be continuously and dynamically updated as more comments are received from a sharing initiator and recipients. It should be understood that the embodiments of the present disclosure are not limited to any specific implementation of the comment function of the sharing message card. Moreover, the embodiments of the present disclosure are not limited to any specific approach of presenting comments in the comment region 1210, e.g., a folded approach, an unfolded approach, etc. may be adopted.

It should be understood that although the examples in the scenario where the target application is an online meeting application are described above in connection with FIG.7 to FIG.12, the embodiments of the present disclosure may also be applied to any other types of target application, and may provide user interfaces, interactive operations, etc. suitable for these target applications. FIG.13 illustrates a flowchart of an exemplary method 1300 for processing a video recording of a target application according to an embodiment.

At 1310, a video recording of the target application may be obtained.

At 1320, multi-modal data of the video recording may be obtained, the multi-modal data comprising at least one of speech transcript, video, image, text and event information.

At 1330, a multi-modal feature of the video recording may be generated based on the multi-modal data, the multi-modal feature comprising at least one of a speech transcript feature, a video feature, an image feature, a text feature and an event feature.

At 1340, target content associated with the video recording may be determined based at least on the multi-modal feature.

In an implementation, the determining target content may comprise: generating a text summary of the video recording.

The generating a text summary may comprise at least one of: generating an extractive summary based at least on the speech transcript feature; and generating an abstractive summary based at least on the speech transcript feature or the extractive summary.

The method 1300 may further comprise: calibrating the extractive summary and/or the abstractive summary with at least the text.

In an implementation, the determining target content may comprise: generating a video summary of the video recording, the video summary comprising at least a part of video frames in the video recording.

The generating a video summary may comprise: obtaining a merged feature based on at least one of the speech transcript feature, the video feature, the image feature and the text feature; and selecting the at least a part of video frames from the video recording based on the merged feature, to form the video summary.

In an implementation, the determining target content may comprise: detecting at least one hot topic in the video recording.

The detecting at least one hot topic may comprise: identifying candidate topics from the speech transcript; and selecting the at least one hot topic from the candidate topics based on at least one of the speech transcript feature, the video feature and the event feature.

The method 1300 may further comprise: extracting at least one hot topic video clip associated with the at least one hot topic from the video recording.

In an implementation, the determining target content may comprise: detecting at least one transcript segment mentioning a target user in the speech transcript based at least on the speech transcript feature; and generating at least one mentioned moment description based on the at least one transcript segment and the event information, and/or extracting at least one mentioned moment video clip from the video recording based on the at least one transcript segment.

In an implementation, the determining target content may comprise: detecting at least one transcript segment containing a task associated with a target user in the speech transcript based at least on the speech transcript feature; and generating at least one task description based on the at least one transcript segment and the event information, and/or extracting at least one task video clip from the video recording based on the at least one transcript segment.

In an implementation, the method 1300 may further comprise: providing a prompt of the target content; and/or presenting the target content.

In an implementation, the method 1300 may further comprise: in response to receiving a request of sharing the target content to at least one recipient, generating a sharing message card associated with the target content; and providing the sharing message card to the at least one recipient.

The sharing message card may have a comment function.

In an implementation, the target application may be at least one of an online meeting application, a video chatting application, a game application, a virtual reality application and a Meta-verse application.

It should be understood that the method 1300 may further comprise any other steps/processes for processing a video recording of a target application according to the embodiments of the present disclosure as described above.

FIG.14 illustrates an exemplary apparatus 1400 for processing a video recording of a target application according to an embodiment.

The apparatus 1400 may include: a video recording obtaining module 1410, for obtaining a video recording of the target application; a multi-modal data obtaining module 1420, for obtaining multi-modal data of the video recording, the multi-modal data comprising at least one of speech transcript, video, image, text and event information; a multi-modal feature generating module 1430, for generating a multi-modal feature of the video recording based on the multi-modal data, the multi-modal feature comprising at least one of a speech transcript feature, a video feature, an image feature, a text feature and an event feature; and a target content determining module 1440, for determining target content associated with the video recording based at least on the multi-modal feature. Moreover, the apparatus 1400 may further comprise any other modules configured for performing any steps/processes of the methods for processing a video recording of a target application according to the embodiments of the present disclosure as described above.

FIG.15 illustrates an exemplary apparatus 1500 for processing a video recording of a target application according to an embodiment.

The apparatus 1500 may comprise at least one processor 1510. The apparatus 1500 may further comprise a memory 1520 connected with the at least one processor 1510. The memory 1520 may store computer-executable instructions that, when executed, cause the at least one processor 1510 to: obtain a video recording of the target application; obtain multi-modal data of the video recording, the multi-modal data comprising at least one of speech transcript, video, image, text and event information; generate a multi-modal feature of the video recording based on the multi-modal data, the multi-modal feature comprising at least one of a speech transcript feature, a video feature, an image feature, a text feature and an event feature; and determine target content associated with the video recording based at least on the multi-modal feature.

In an implementation, the determining target content may comprise at least one of: generating a text summary of the video recording; generating a video summary of the video recording, the video summary comprising at least a part of video frames in the video recording; and detecting at least one hot topic in the video recording.

In an implementation, the determining target content may comprise: detecting at least one

transcript segment mentioning a target user in the speech transcript based at least on the speech transcript feature; and generating at least one mentioned moment description based on the at least one transcript segment and the event information, and/or extracting at least one mentioned moment video clip from the video recording based on the at least one transcript segment.

5 In an implementation, the determining target content may comprise: detecting at least one transcript segment containing a task associated with a target user in the speech transcript based at least on the speech transcript feature; and generating at least one task description based on the at least one transcript segment and the event information, and/or extracting at least one task video clip from the video recording based on the at least one transcript segment.

10 Moreover, the at least one processor 1510 may be further configured for performing any other steps/processes of the methods for processing a video recording of a target application according to the embodiments of the present disclosure as described above.

The embodiments of the present disclosure propose a computer program product for processing a video recording of a target application. The computer program product may comprise a computer
15 program that is executed by at least one processor for: obtaining a video recording of the target application; obtaining multi-modal data of the video recording, the multi-modal data comprising at least one of speech transcript, video, image, text and event information; generating a multi-modal feature of the video recording based on the multi-modal data, the multi-modal feature comprising at least one of a speech transcript feature, a video feature, an image feature, a text
20 feature and an event feature; and determining target content associated with the video recording based at least on the multi-modal feature. The computer program may be further executed by the at least one processor for performing any other steps/processes of the methods for processing a video recording of a target application according to the embodiments of the present disclosure as described above.

25 The embodiments of the present disclosure may be embodied in a non-transitory computer-readable medium. The non-transitory computer readable medium may comprise instructions that, when executed, cause one or more processors to perform any steps/processes of the methods for processing a video recording of a target application according to the embodiments of the present disclosure as described above.

30 It should be appreciated that all the operations in the methods described above are merely exemplary, and the present disclosure is not limited to any operations in the methods or sequence orders of these operations, and should cover all other equivalents under the same or similar concepts.

Moreover, the articles "a" and "an" as used in this description and appended claims, unless
35 otherwise specified or clear from the context that they are for the singular form, should generally

be interpreted as meaning "one" or "one or more."

It should also be appreciated that all the modules in the apparatuses described above may be implemented in various approaches. These modules may be implemented as hardware, software, or a combination thereof. Moreover, any of these modules may be further functionally divided into sub-modules or combined together.

Processors have been described in connection with various apparatuses and methods. These processors may be implemented using electronic hardware, computer software, or any combination thereof. Whether such processors are implemented as hardware or software will depend upon the particular application and overall design constraints imposed on the system. By way of example, a processor, any portion of a processor, or any combination of processors presented in this disclosure may be implemented as a microprocessor, a microcontroller, a digital signal processor (DSP), a field programmable gate array (FPGA), a programmable logic device (PLD), state machine, gate logic, discrete hardware circuitry, and other suitable processing components configured to perform the various functions described in this disclosure. The functions of a processor, any portion of a processor, or any combination of processors presented in this disclosure may be implemented as software executed by a microprocessor, a microcontroller, a DSP, or other suitable platforms.

Software shall be construed broadly to mean instructions, instruction sets, code, code segments, program code, programs, subprograms, software modules, applications, software applications, software packages, routines, subroutines, objects, threads of execution, procedures, functions, etc. The software may reside on a computer-readable medium. A computer-readable medium may include, by way of example, memory such as a magnetic storage device (e.g., hard disk, floppy disk, magnetic strip), an optical disk, a smart card, a flash memory device, random access memory (RAM), read only memory (ROM), programmable ROM (PROM), erasable PROM (EPROM), electrically erasable PROM (EEPROM), a register, or a removable disk. Although a memory is shown as being separate from the processor in various aspects presented in this disclosure, a memory may also be internal to the processor (e.g., a cache or a register).

The previous description is provided to enable any person skilled in the art to practice the various aspects described herein. Various modifications to these aspects will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other aspects. Thus, the claims are not intended to be limited to the aspects shown herein. All structural and functional equivalents to the elements of the various aspects described throughout the present disclosure that are known or later come to be known to those of ordinary skilled in the art are intended to be encompassed by the claims.

CLAIMS

1. A method for processing a video recording of a target application, comprising:
obtaining a video recording of the target application;
obtaining multi-modal data of the video recording, the multi-modal data comprising at least one of speech transcript, video, image, text and event information;
generating a multi-modal feature of the video recording based on the multi-modal data, the multi-modal feature comprising at least one of a speech transcript feature, a video feature, an image feature, a text feature and an event feature; and
determining target content associated with the video recording based at least on the multi-modal feature.
2. The method of claim 1, wherein the determining target content comprises:
generating a text summary of the video recording.
3. The method of claim 2, wherein the generating a text summary comprises at least one of:
generating an extractive summary based at least on the speech transcript feature; and
generating an abstractive summary based at least on the speech transcript feature or the extractive summary.
4. The method of claim 3, further comprising:
calibrating the extractive summary and/or the abstractive summary with at least the text.
5. The method of claim 1, wherein the determining target content comprises:
generating a video summary of the video recording, the video summary comprising at least a part of video frames in the video recording.
6. The method of claim 5, wherein the generating a video summary comprises:
obtaining a merged feature based on at least one of the speech transcript feature, the video feature, the image feature and the text feature; and
selecting the at least a part of video frames from the video recording based on the merged feature, to form the video summary.
7. The method of claim 1, wherein the determining target content comprises:
detecting at least one hot topic in the video recording.
8. The method of claim 7, wherein the detecting at least one hot topic comprises:
identifying candidate topics from the speech transcript; and
selecting the at least one hot topic based on at least one of the speech transcript feature, the video feature and the event feature.
9. The method of claim 7, further comprising:
extracting at least one hot topic video clip associated with the at least one hot topic from the video recording.

10. The method of claim 1, wherein the determining target content comprises:
detecting at least one transcript segment mentioning a target user in the speech transcript based at least on the speech transcript feature; and
generating at least one mentioned moment description based on the at least one transcript segment and the event information, and/or extracting at least one mentioned moment video clip from the video recording based on the at least one transcript segment.
11. The method of claim 1, wherein the determining target content comprises:
detecting at least one transcript segment containing a task associated with a target user in the speech transcript based at least on the speech transcript feature; and
generating at least one task description based on the at least one transcript segment and the event information, and/or extracting at least one task video clip from the video recording based on the at least one transcript segment.
12. The method of claim 1, further comprising:
providing a prompt of the target content; and/or
presenting the target content.
13. The method of claim 1, further comprising:
in response to receiving a request of sharing the target content to at least one recipient, generating a sharing message card associated with the target content; and
providing the sharing message card to the at least one recipient.
14. An apparatus for processing a video recording of a target application, comprising:
at least one processor; and
a memory storing computer-executable instructions that, when executed, cause the at least one processor to:
obtain a video recording of the target application,
obtain multi-modal data of the video recording, the multi-modal data comprising at least one of speech transcript, video, image, text and event information,
generate a multi-modal feature of the video recording based on the multi-modal data, the multi-modal feature comprising at least one of a speech transcript feature, a video feature, an image feature, a text feature and an event feature, and
determine target content associated with the video recording based at least on the multi-modal feature.
15. A computer program product for processing a video recording of a target application, comprising a computer program that is executed by at least one processor for:
obtaining a video recording of the target application;
obtaining multi-modal data of the video recording, the multi-modal data comprising at least

one of speech transcript, video, image, text and event information;

generating a multi-modal feature of the video recording based on the multi-modal data, the multi-modal feature comprising at least one of a speech transcript feature, a video feature, an image feature, a text feature and an event feature; and

determining target content associated with the video recording based at least on the multi-modal feature.

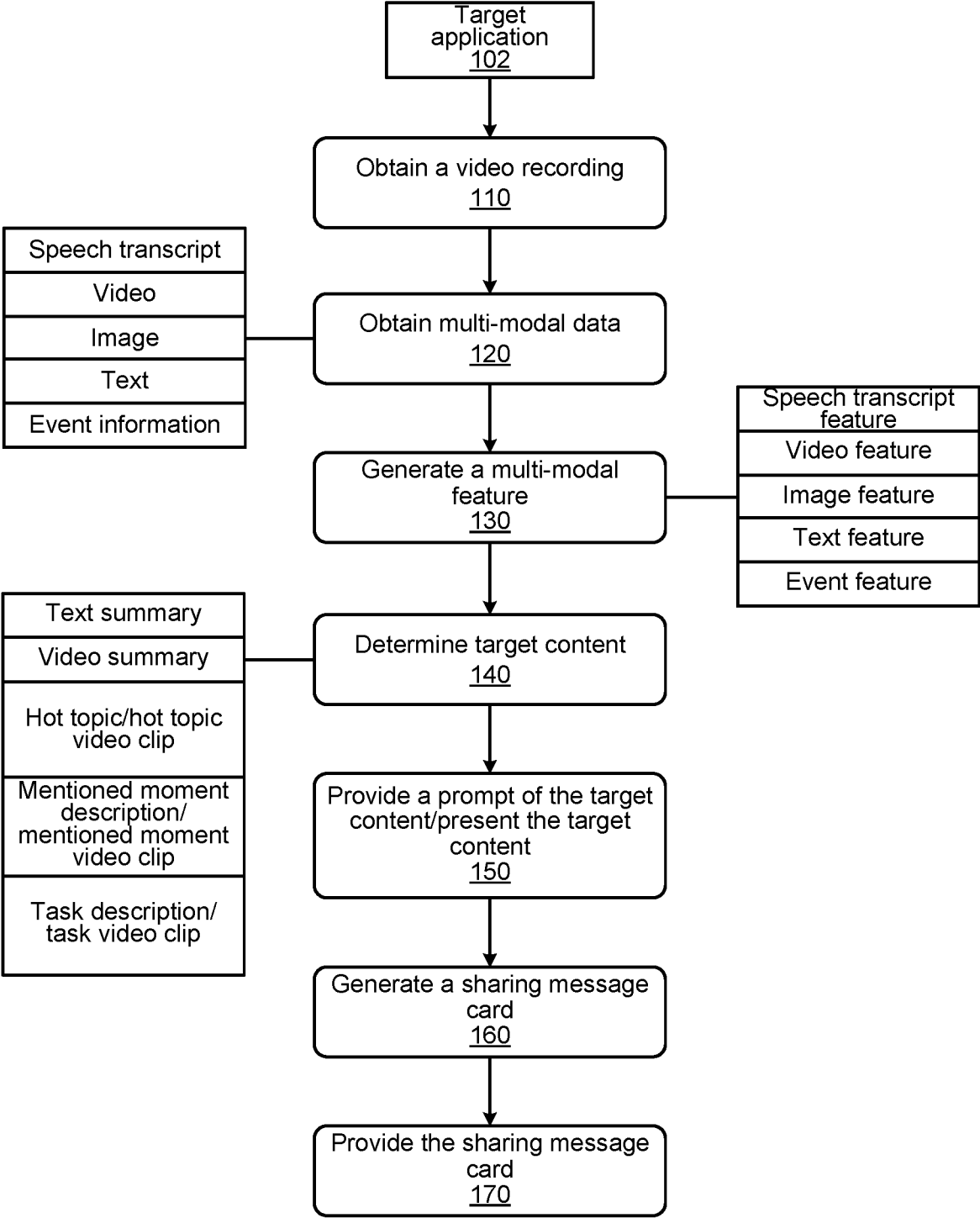


FIG 1

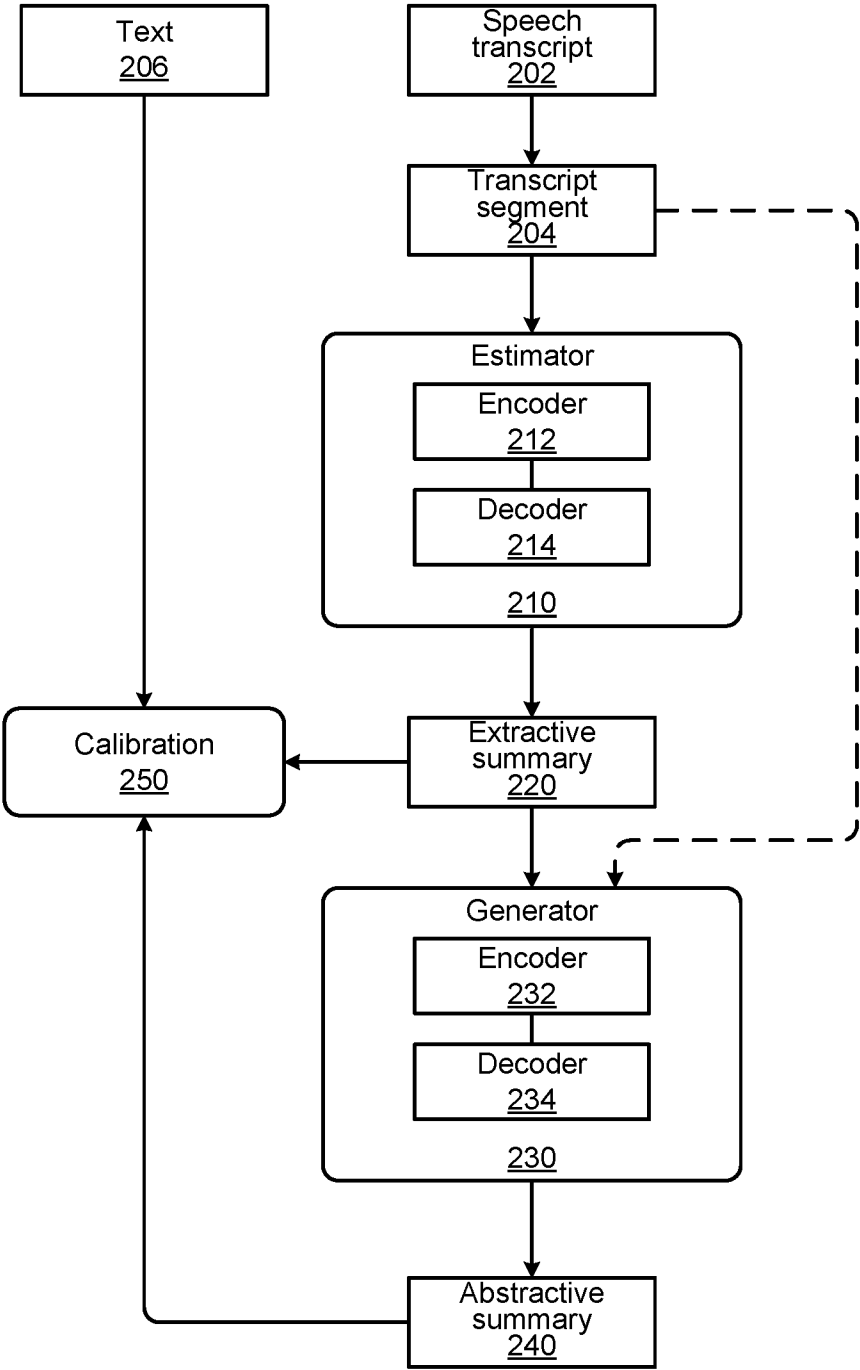


FIG 2

3/10

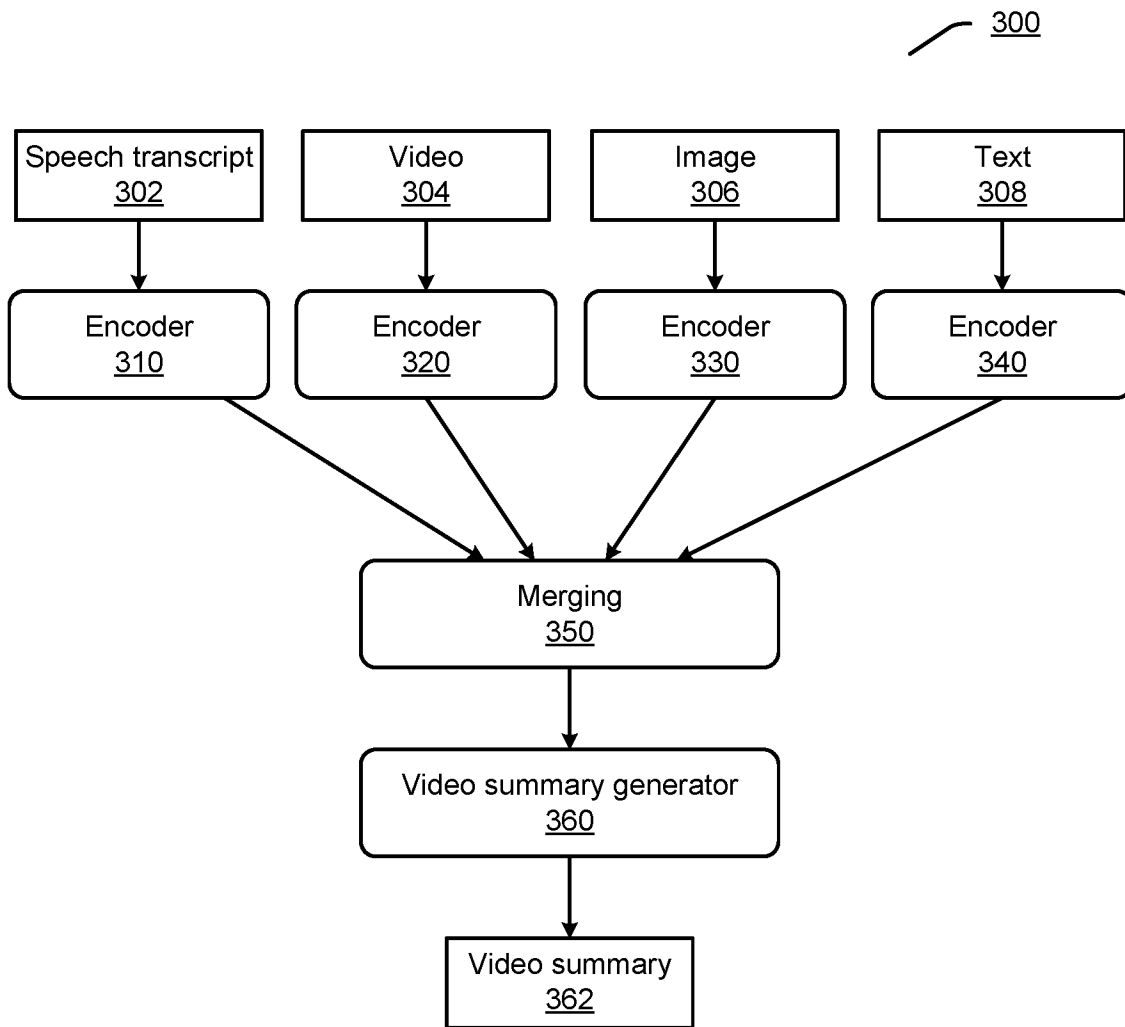


FIG 3

4/10

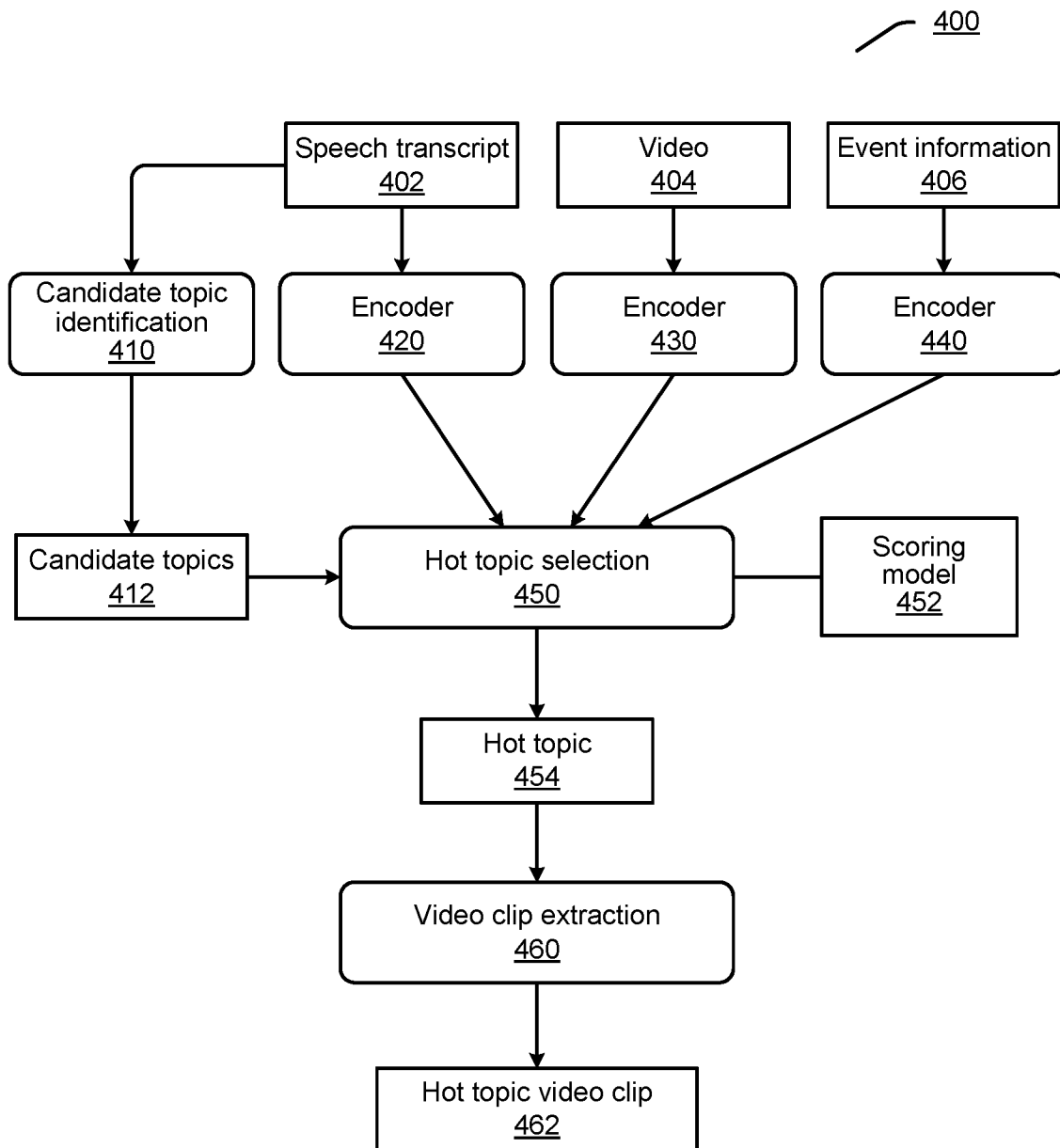


FIG 4

5/10

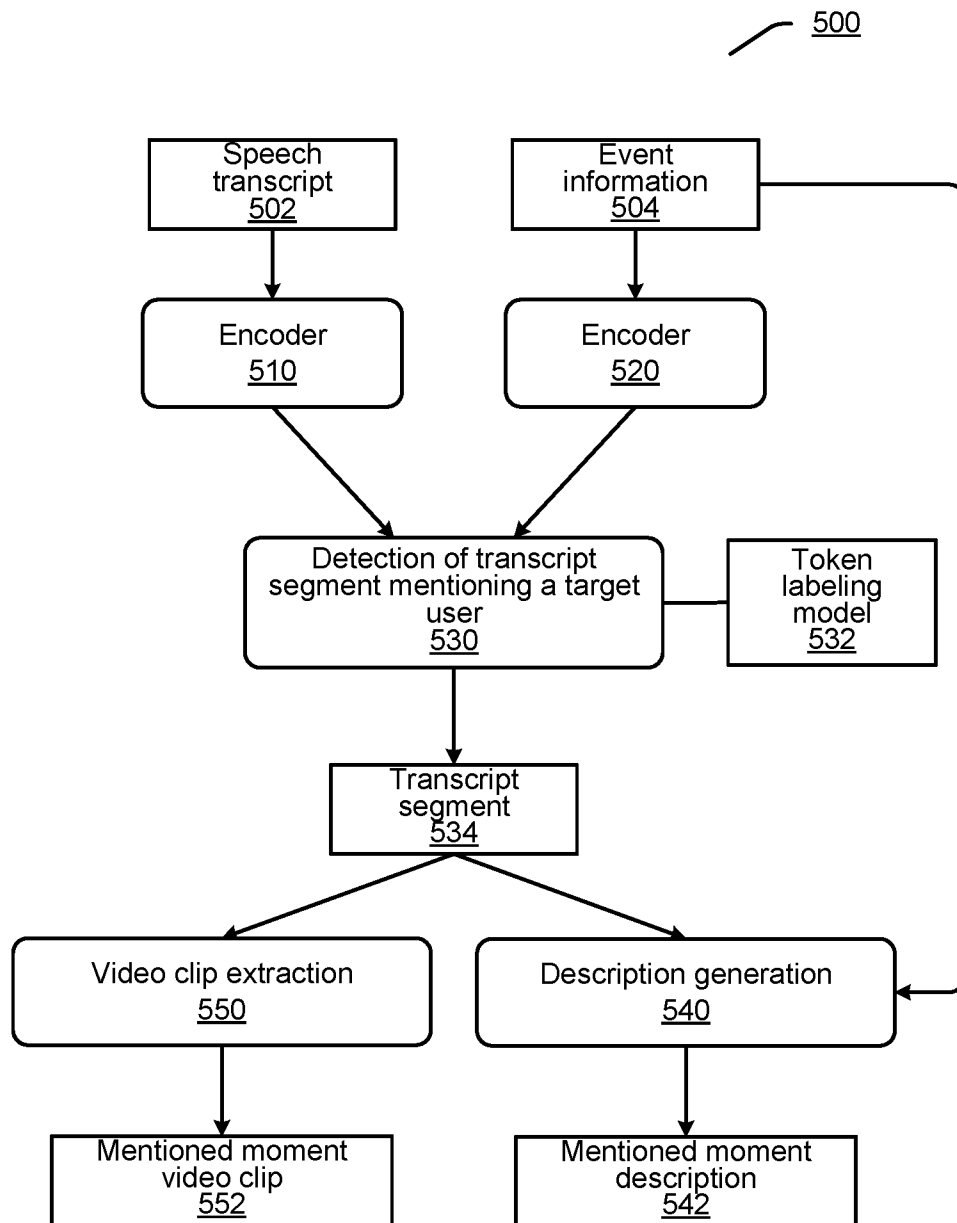


FIG 5

6/10

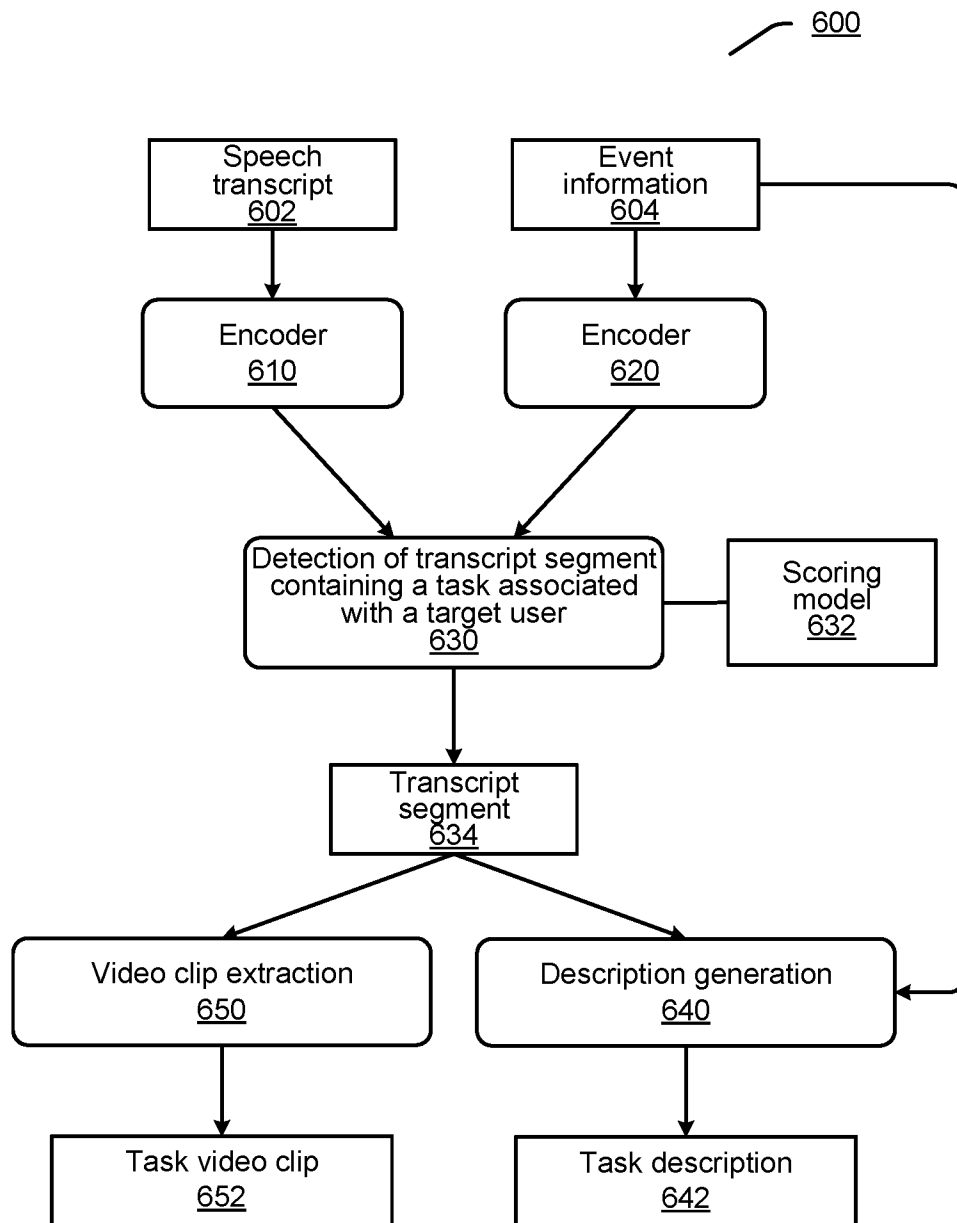


FIG 6

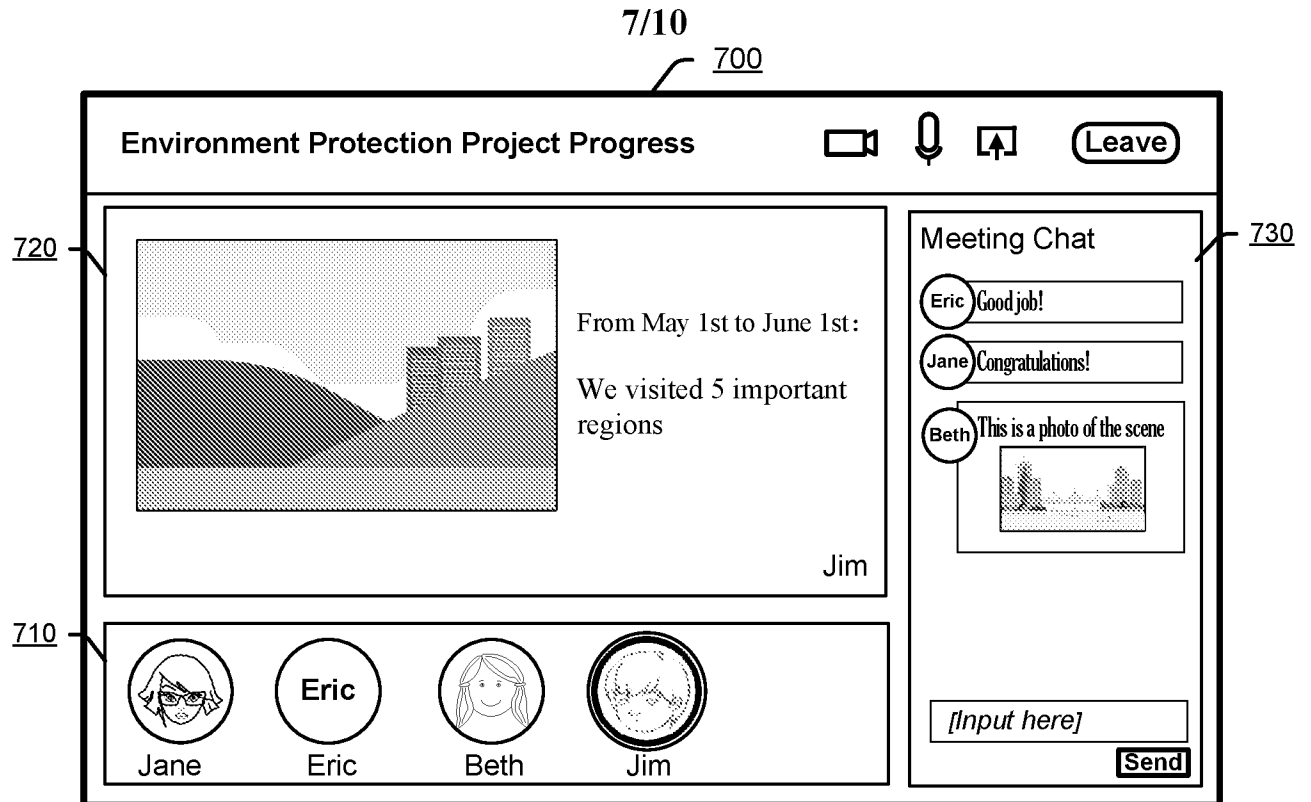


FIG 7

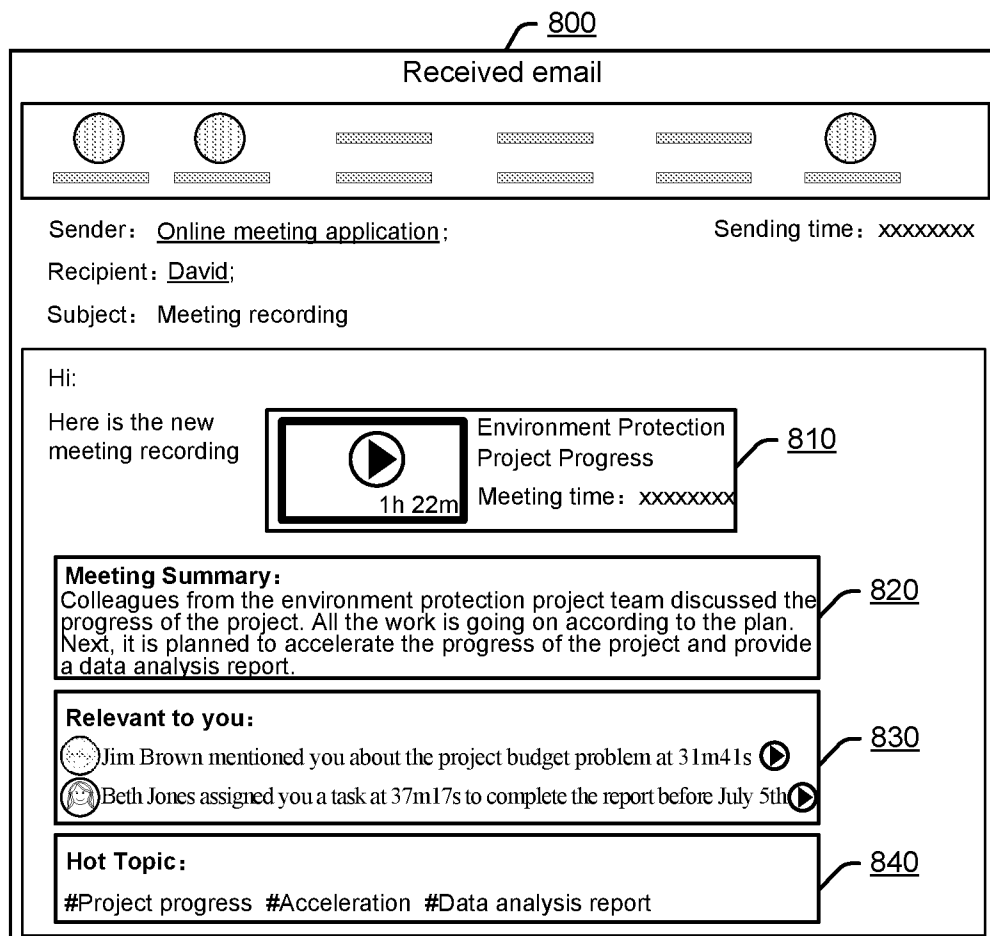


FIG 8

8/10

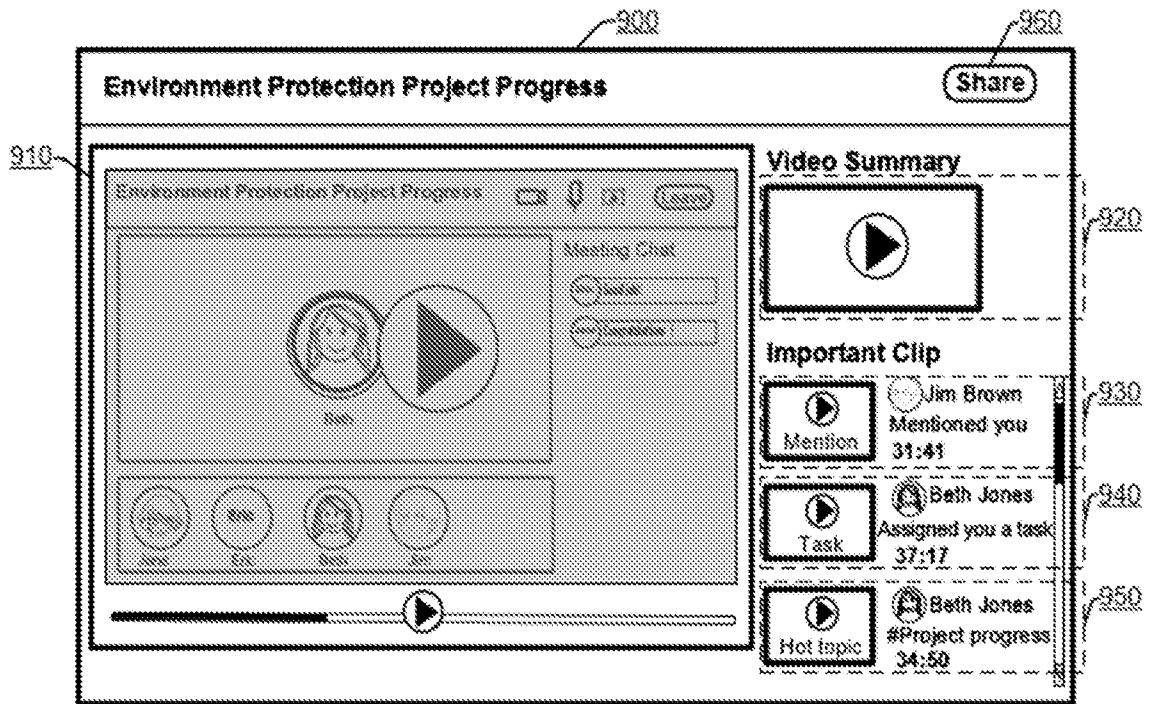


FIG 9

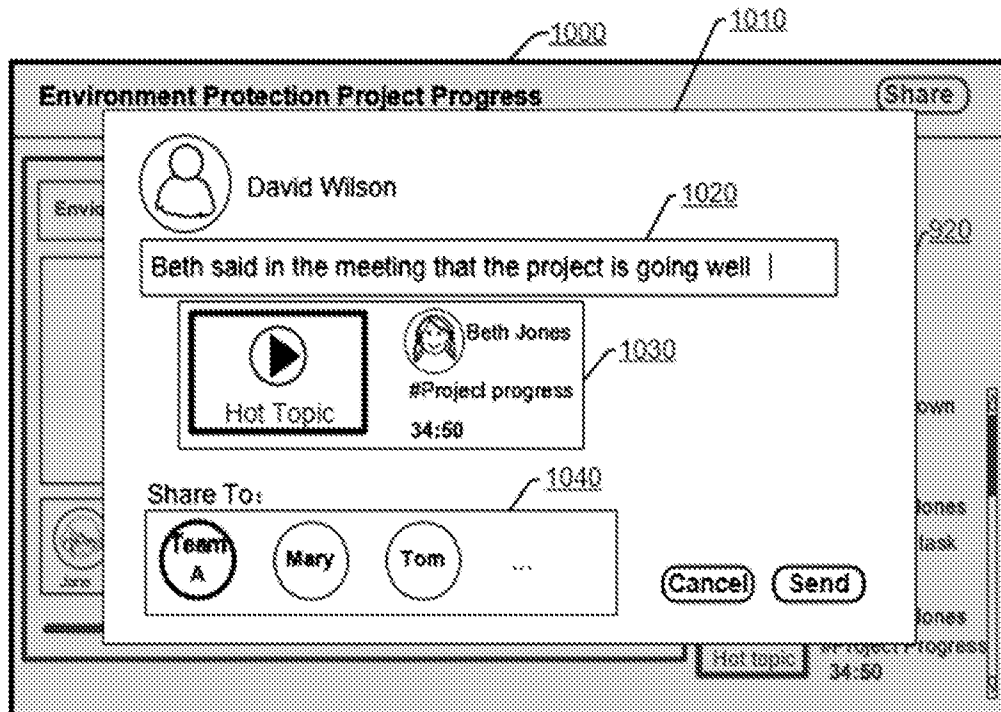


FIG 10

9/10

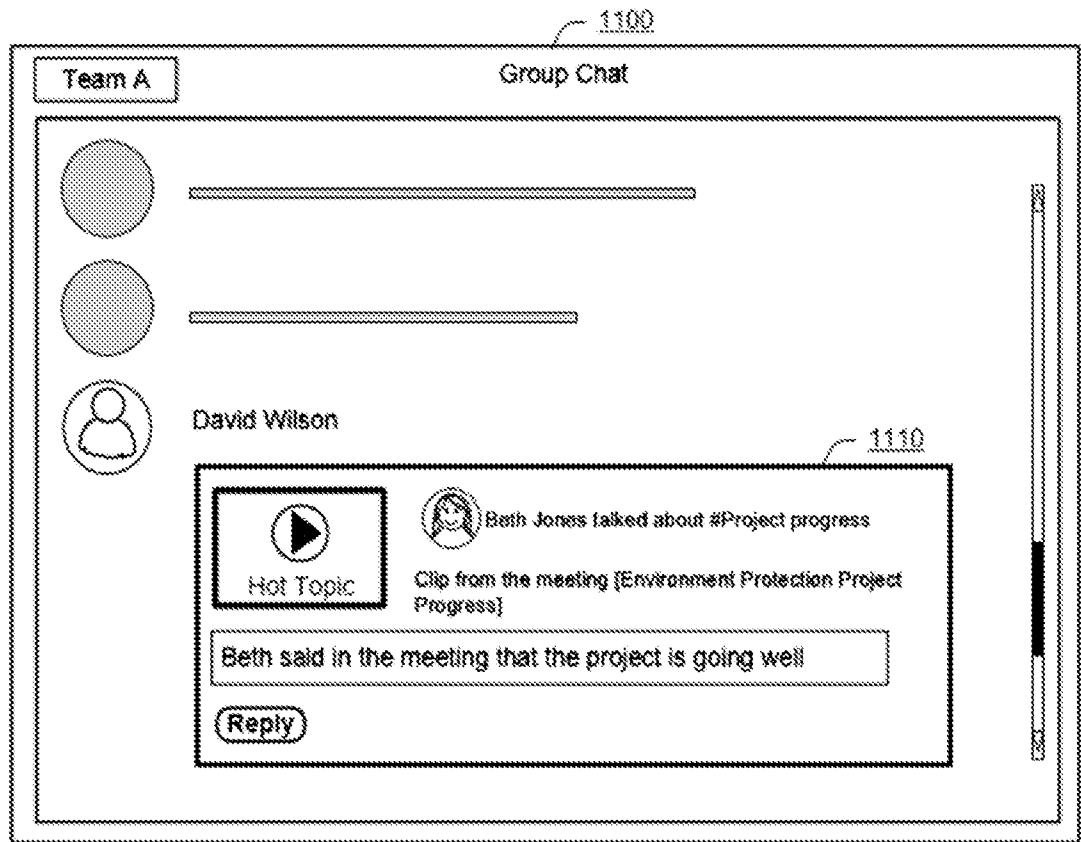


FIG 11

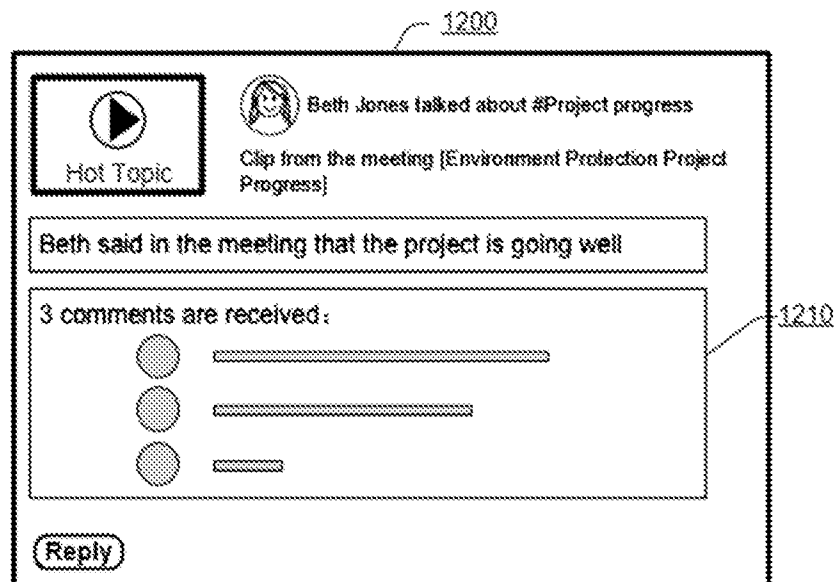
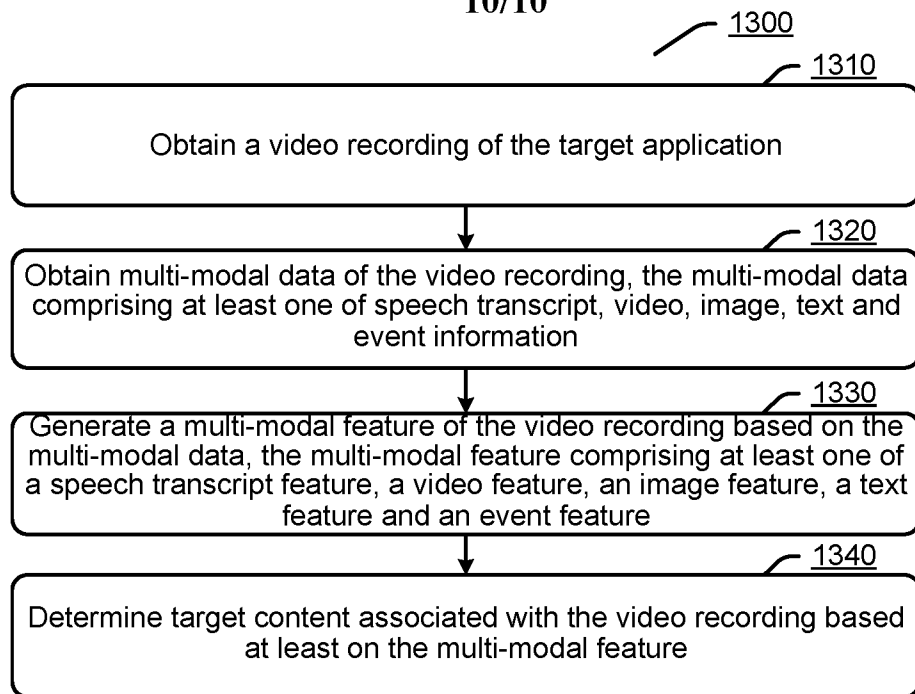
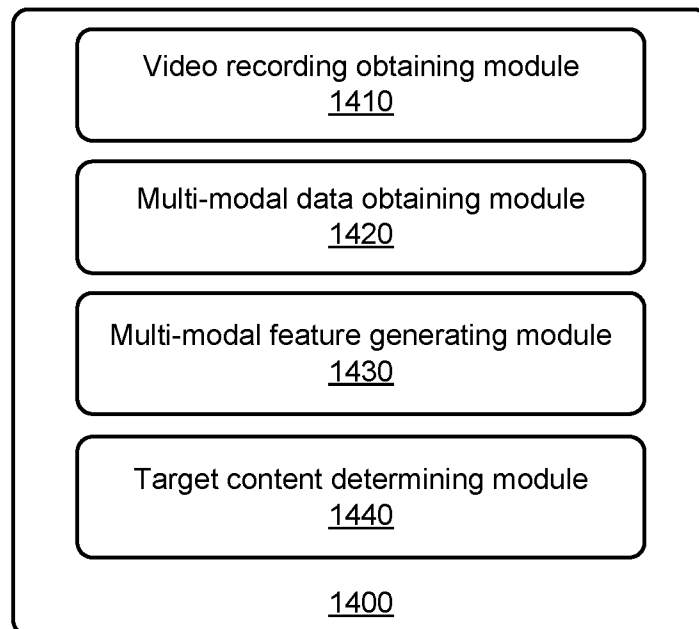
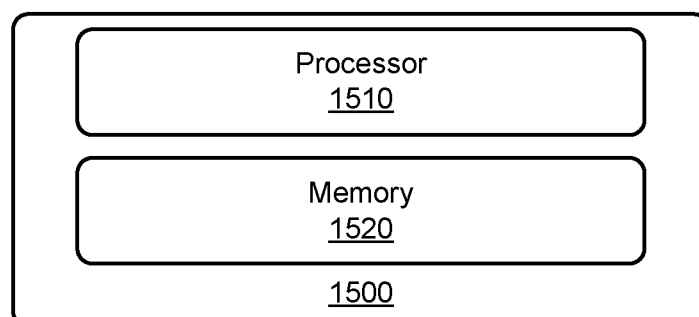


FIG 12

10/10

**FIG 13****FIG 14****FIG 15**

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2023/019039

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06F16/738

ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2013/325972 A1 (BOSTON JEFFREY S [US] ET AL) 5 December 2013 (2013-12-05) abstract paragraphs [0032], [0033], [0046], [0068] - [0070]; figure 4A -----	1-15
X	US 2021/383127 A1 (KIKIN-GIL EREZ [US] ET AL) 9 December 2021 (2021-12-09) abstract paragraphs [0003] - [0005], [0021], [0022], [0046] - [0050]; figure 2 -----	1-15
X	US 2019/057258 A1 (SIMSKE STEVEN J [US] ET AL) 21 February 2019 (2019-02-21) abstract paragraphs [0002] - [0011], [0037] - [0422]; figure 2 -----	1-15



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

20 June 2023

Date of mailing of the international search report

27/06/2023

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2

NL - 2280 HV Rijswijk

Tel. (+31-70) 340-2040,

Fax: (+31-70) 340-3016

Authorized officer

Siódsmok, Wojciech

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2023/019039

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2013325972 A1	05-12-2013	NONE	
US 2021383127 A1	09-12-2021	CN 115668371 A	31-01-2023
		EP 4162488 A1	12-04-2023
		US 2021383127 A1	09-12-2021
		WO 2021247156 A1	09-12-2021
US 2019057258 A1	21-02-2019	CN 108353213 A	31-07-2018
		EP 3369252 A1	05-09-2018
		US 2019057258 A1	21-02-2019
		WO 2017074448 A1	04-05-2017