CLAIMS:

1. A comment generation system comprising:

a processor and a memory operatively coupled to the processor; and

a comment generation program stored in the memory and executed by the processor to be configured to:

receive an input video in a social networking system;

generate at least one predicted comment corresponding to the input video by projecting video-frame embeddings of the input video and text embeddings associated with a user profile of a target user into a shared cross-modal embedding space using a cross-modal embedding model trained on a dataset of comment and image pairs;

compute, within the shared cross-modal embedding space, probability scores for the comment and image pairs;

select the at least one predicted comment from the comment and image pairs based on the probability scores; and

present the at least one predicted comment to the target user.

2. The comment generation system of claim 1, wherein the at least one predicted comment is selected from a plurality of comment candidates.

3. The comment generation system of claim 1, wherein the at least one predicted comment is generated based on one or more comments made by the target user which precede the input video.

4. The comment generation system of claim 1, wherein the comment generation program is further configured to:

generate video tags corresponding to the input video; and

generate the at least one predicted comment based on the video tags.

5. The comment generation system of claim 4, wherein the video tags are generated based on at least one selected from the group of themes, genres, objects, and actions.

6. The comment generation system of claim 4, wherein the video tags are generated based on readable text detected within the video frames of the input video.

7. The comment generation system of claim 6, wherein the input video is labeled based on transcribed audio of the input video generated based on a speech recognition model.

8. The comment generation system of claim 1, wherein the comment generation program is further configured to:

extract embeddings of the input video; and

generate the at least one predicted comment based on the extracted embeddings.

9. The comment generation system of claim 1, wherein the at least one predicted comment is at least one selected from the group of text, an emoji, and an image.

10. A method comprising:

receiving an input video in a social networking system;

generating at least one predicted comment corresponding to the input video by projecting video-frame embeddings of the input video and text embeddings associated with a user profile of a target user into a shared cross-modal embedding space using a cross-modal embedding model trained on a dataset of comment and image pairs;

computing, within the shared cross-modal embedding space, probability scores for the comment and image pairs;

selecting the at least one predicted comment from the comment and image pairs based on the probability scores; and

presenting the at least one predicted comment to the target user.

11. The method of claim 10, wherein the at least one predicted comment is selected from a plurality of comment candidates.

12. The method of claim 10, wherein the at least one predicted comment is generated based on one or more comments made by the target user which precede the input video.

13. The method of claim 10, further comprising:

generating video tags corresponding to the input video; and

generating the at least one predicted comment based on the video tags.

14. The method of claim 13, wherein the video tags are generated based on at least one selected from the group of themes, genres, objects, and actions.

15. The method of claim 13, wherein the video tags are generated based on readable text detected within the video frames of the input video.

16. The method of claim 15, wherein the input video is labeled based on transcribed audio of the input video generated based on a speech recognition model.

17. The method of claim 10, further comprising:

extracting embeddings of the input video; and

generating the at least one predicted comment based on the extracted embeddings.

18. A comment generation system comprising:

a processor and a memory operatively coupled to the processor; and

a comment generation program stored in the memory and executed by the processor to be configured to:

receive an input video in a social networking system;

generate a plurality of comment candidates corresponding to the input video by projecting video-frame embeddings of the input video and text embeddings associated with a user profile of a target user into a shared cross-modal embedding space using a cross-modal embedding model trained on a dataset of comment and image pair;

compute, within the shared cross-modal embedding space, probability scores for the comment and image pairs;

select the plurality of comment candidates from the comment and image pairs based on the probability scores;

present the plurality of comment candidates to the target user;

receive a selection of one of the plurality of comment candidates from the target user; and

post the selected comment candidate as a reply to the input video.