CLAIMS

1. A method comprising:

receiving visual training data at a visual encoder comprising a first machine learning (ML) model;

identifying, using the first ML model executed using one or more processor, data corresponding to a visual object in the visual training data;

receiving audio training data synchronized to the visual training data at an audio forwarding regularizer comprising a second ML model, wherein the audio training data comprises a visually relevant sound and a visually irrelevant sound that are both synchronized to a same frame in the visual training data that includes the visual object, wherein the visually relevant sound corresponds to the visual object but the visually irrelevant sound is generated by an audio source that is not visible in the same frame;

filtering out data corresponding to the visually relevant sound from an output of the second ML model using an information bottleneck such that only data corresponding to the visually irrelevant sound is primarily output from the information bottleneck; and

training a third ML model downstream from the first and second ML models using the data corresponding to the visual object and data corresponding to the visually irrelevant sound received from the information bottleneck.

2. The method of claim 1, further comprising, after training the third ML model:

performing an execution stage, the execution stage comprising:

receiving a silent video at the first ML model;

identifying, using the first ML model, data corresponding to a second visual object in the silent video;

generating, using the third ML model, a visually relevant sound to synchronize to at least one video frame of the silent video containing the second visual object, wherein data corresponding to the second visual object is inputted to the third ML model; and

generating a media presentation based on the synchronized visually relevant sound and video frames in the silent video.

3. The method of claim 2, wherein the second ML model in the audio forwarding regularizer is unused during the execution stage.

4. The method of claim 1, wherein the information bottleneck comprises limiting a dimension of an output of the second ML model, wherein limiting the dimension prevents data corresponding to the visually relevant sound from reaching the third ML model.

5. The method of claim 1, wherein the second ML model comprises a sequence-to-sequence ML model.

6. The method of claim 5, wherein the sequence-to-sequence ML model outputs a bottlenecked audio frame based on the audio training data and the information bottleneck, the method further comprising:

replicating the bottlenecked audio frame according to a number of T time periods in the visual training data; and

transmitting the replicated bottlenecked audio frames to the third ML model.

7. The method of claim 6, further comprising:

predicting, using the third ML model, the visually relevant and irrelevant sounds in the audio training data based on receiving the replicated bottlenecked audio frames and the data corresponding to a visual object.

8. A computer program product, comprising:

a computer-readable storage medium having computer-readable program code embodied therewith, the computer-readable program code executable by one or more computer processors to perform an operation, the operation comprising:

receiving visual training data at a visual encoder comprising a first ML model;

identifying, using the first ML model, data corresponding to a visual object in the visual training data;

receiving audio training data synchronized to the visual training data at an audio forwarding regularizer comprising a second ML model, wherein the audio training data comprises a visually relevant sound and a visually irrelevant sound that are both synchronized to a same frame in the visual training data that includes the visual object, wherein the visually relevant sound corresponds to the visual object but the visually irrelevant sound is generated by an audio source that is not visible in the same frame;

filtering out data corresponding to the visually relevant sound from an output of the second ML model using an information bottleneck such that only data corresponding to the visually irrelevant sound is primarily output from the information bottleneck; and

training a third ML model downstream from the first and second ML models using the data corresponding to the visual object and data corresponding to the visually irrelevant sound received from the information bottleneck.

9. The computer program product of claim 8, wherein the operation further comprises, after training the third ML model:

performing an execution stage, the execution stage comprising:

receiving a silent video at the first ML model;

identifying, using the first ML model, data corresponding to a second visual object in the silent video;

generating, using the third ML model, a visually relevant sound to synchronize to at least one video frame of the silent video containing the second visual object, wherein data corresponding to the second visual object is inputted to the third ML model; and

generating a media presentation based on the synchronized visually relevant sound and video frames in the silent video.

10. The computer program product of claim 9, wherein the second ML model in the audio forwarding regularizer is unused during the execution stage.

11. The computer program product of claim 8, wherein the information bottleneck comprises limiting a dimension of an output of the second ML model, wherein limiting the dimension prevents data corresponding to the visually relevant sound from reaching the third ML model.

12. The computer program product of claim 8, wherein the second ML model comprises a sequence-to-sequence ML model.

13. The computer program product of claim 12, wherein the sequence-to-sequence ML model outputs a bottlenecked audio frame based on the audio training data and the information bottleneck, the operation further comprising:

replicating the bottlenecked audio frame according to a number of T time periods in the visual training data; and

transmitting the replicated bottlenecked audio frames to the third ML model.

14. The computer program product of claim 13, wherein the operation further comprises:

predicting, using the third ML model, the visually relevant and irrelevant sounds in the audio training data based on receiving the replicated bottlenecked audio frames and the data corresponding to a visual object.

15. A system, comprising:

a processor; and

a memory comprising a program, which when executed by the processor, performs an operation, the operation comprising:

receiving visual training data at a visual encoder comprising a first ML model;

identifying, using the first ML model, data corresponding to a visual object in the visual training data;

receiving audio training data synchronized to the visual training data at an audio forwarding regularizer comprising a second ML model, wherein the audio training data comprises a visually relevant sound and a visually irrelevant sound that are both synchronized to a same frame in the visual training data that includes the visual object, wherein the visually relevant sound corresponds to the visual object but the visually irrelevant sound is generated by an audio source that is not visible in the same frame;

filtering out data corresponding to the visually relevant sound from an output of the second ML model using an information bottleneck such that only data corresponding to the visually irrelevant sound is primarily output from the information bottleneck; and

training a third ML model downstream from the first and second ML models using the data corresponding to the visual object and data corresponding to the visually irrelevant sound received from the information bottleneck.

16. The system of claim 15, wherein the operation further comprises, after training the third ML model:

performing an execution stage, the execution stage comprising:

receiving a silent video at the first ML model;

identifying, using the first ML model, data corresponding to a second visual object in the silent video;

generating, using the third ML model, a visually relevant sound to synchronize to at least one video frame of the silent video containing the second visual object, wherein data corresponding to the second visual object is inputted to the third ML model; and

generating a media presentation based on the synchronized visually relevant sound and video frames in the silent video.

17. The system of claim 16, wherein the second ML model in the audio forwarding regularizer is unused during the execution stage.

18. The system of claim 15, wherein the information bottleneck comprises limiting a dimension of an output of the second ML model, wherein limiting the dimension prevents data corresponding to the visually relevant sound from reaching the third ML model.

19. The system of claim 15, wherein the second ML model comprises a sequence-to-sequence ML model and wherein the sequence-to-sequence ML model outputs a bottlenecked audio frame based on the audio training data and the information bottleneck, the operation further comprising:

replicating the bottlenecked audio frame according to a number of T time periods in the visual training data; and

transmitting the replicated bottlenecked audio frames to the third ML model.

20. The system of claim 19, wherein the operation further comprises:

predicting, using the third ML model, the visually relevant and irrelevant sounds in the audio training data based on receiving the replicated bottlenecked audio frames and the data corresponding to a visual object.