WHAT IS CLAIMED IS:

1. A method comprising:

receiving a media item;

generating, using a media encoder of a trained language-media model, an encoded media item from the media item, wherein the encoded media item is in a vector format;

inputting the encoded media item to a machine learning model, wherein the machine learning model is trained using media items having known classifications for corresponding encoded media items; and

determining, using the machine learning model, a classification for the media item.

2. The method of claim 1, further comprising:

generating, by the machine learning model, an encoded prompt output, the encoded prompt output in vector format; and

generating, using a text decoder, a prompt using the encoded prompt output, the prompt describing features of the media item.

3. The method of claim 1, wherein the classification is extracted from the generated prompt.

4. The method of claim 1, wherein training the machine learning model further comprises:

receiving training media, each of the training media having a known classification of a plurality of known classifications;

processing, using the media encoder, each of the training media to obtain training media encoding vectors;

processing, using the machine learning model, the training media encoding vectors to obtain a predicted classification for each of the training media;

for each of the training media, comparing the predicted classification to the known classification; and

updating parameters of the machine learning model based on the comparison.

5. The method of claim 4, wherein training the machine learning model further comprises:

receiving a training text for each of the training media, the training text including the known classification;

processing, using a text encoder, each training text to obtain training text encoding vectors,

wherein processing, using the machine learning model, the training media encoding vectors to obtain the predicted classification comprises:

processing, using the machine learning model, the training media encoding vectors and the training text encoding vectors to determine matching text-media pairs, wherein the training text of each of the text-media pairs includes the predicted classification of a corresponding training media item.

6. The method of claim 1, wherein the machine learning model includes a neural network.

7. The method of claim 1, wherein the machine learning model provides a probability for each of the known classifications and selects the classification with a highest probability.

8. The method of claim 1, wherein determining a classification for the media item further comprises:

retrieving one or more reference encoded prompt vectors, each corresponding to a known classification;

determining a mapped vector by mapping from the encoded media item to the known classification;

comparing the mapped vector to the one or more reference encoded prompt vectors to identify a closest reference encoded prompt vector; and

determining the classification based on the closest reference encoded prompt vector.

9. A system comprising:

one or more storage media storing computer-executable instructions; and

one or more processors configured to access the one or more storage media and execute the computer-executable instructions to perform the method of any one of claims 1-8.

10. A method comprising:

receiving a media item;

generating, using media encoder of a language-media model, an encoded media item from the media item, wherein the encoded media item is in a vector format within an encoded media space;

inputting the encoded media item into a machine learning model that is trained to select an encoded text vector corresponding to a correct classification of the media item;

determining, by the machine learning model, an encoded prompt output, the encoded prompt output in vector format; and

generating, using a text decoder, a prompt using the encoded prompt output, the prompt describing features of the media item.

11. The method of claim 10, wherein the media item is an image, an audio file, or a video.

12. The method of claim 10, wherein the prompt includes co-occurring features of the media item.

13. The method of claim 10, wherein training the text decoder comprises:

receiving a training prompt;

generating, using a text encoder of the language-media model, an encoded prompt from the training prompt, wherein the encoded prompt is in vector format;

processing, by the text decoder, the encoded prompt to obtain a predicted prompt;

comparing the predicted prompt to the training prompt; and

updating parameters of the text decoder based on the comparison.

14. The method of claim 10, wherein the training involves determining a closest encoded text vector.

15. A system comprising:

one or more storage media storing computer-executable instructions; and

one or more processors configured to access the one or more storage media and execute the computer-executable instructions to perform the method of any one of claims 10-14.