CLAIMS:

1. A computer-implemented method for creating content, the method comprising:

receiving text, wherein the text is associated with a video to be created by at least one user;

receiving a style selection other than the text associated with the video to be created, wherein the style selection comprises input by selecting an icon corresponding to the desired style on an interface; generating at least one image based at least in part on the text and the style selection using at least one machine learning model; and

generating the video based at least in part on the at least one image by overlaying additional content onto the at least one image.

2. The method of claim 1, wherein the style selection indicates at least one of a color, a texture, or an artistic style.

3. The method of claim 1, wherein the at least one machine learning model comprises:

a first machine learning model configured to receive the text and generate at least one initial image based at least in part on the text; and

a second machine learning model configured to receive a plurality of images and generate a plurality of text captions corresponding to the plurality of images,

wherein the first machine learning model is trained on a dataset comprising the plurality of images and the plurality of corresponding text captions.

4. The method of claim 3, wherein the at least one machine learning model further comprises:

a third machine learning model configured to receive the text and the at least one initial image, the third machine learning model further configured to generate the at least one image based on the text and the at least one initial image, wherein the at least one image has a higher resolution than the at least one initial image.

5. The method of claim 1, wherein the text comprises input from the at least one user, and wherein the additional content comprises at least one portion of a live camera feed.

6. The method of claim 1, wherein the text comprises a plurality of song lyrics, and wherein the video comprises a plurality of images corresponding to the plurality of song lyrics.

7. The method of claim 1, wherein the at least one image comprises different images generated based at least in part on a same text input by different users.

8. The method of claim 7, wherein the video comprises content overlaid on at least one subset of the plurality of different images.

9. A system, comprising:

at least one processor in communication with at least one memory, the at least one memory comprising computer-readable instructions that upon execution by the at least one processor configure the system to perform operations comprising:

receiving text, wherein the text is associated with a video to be created by at least one user;

receiving a style selection other than the text associated with the video to be created, wherein the style selection comprises input by selecting an icon corresponding to the desired style on an interface;

generating at least one image based at least in part on the text and the style selection using at least one machine learning model; and

generating the video based at least in part on the at least one image by overlaying additional content onto the at least one image.

10. The system of claim 9, wherein the style selection indicates at least one of a color, a texture, or an artistic style.

11. The system of claim 9, wherein the at least one machine learning model comprises:

a first machine learning model configured to receive the text and generate at least one initial image based at least in part on the text; and

a second machine learning model configured to receive a plurality of images and generate a plurality of text captions corresponding to the plurality of images,

wherein the first machine learning model is trained on a dataset comprising the plurality of images and the plurality of corresponding text captions.

12. The system of claim 11, wherein the at least one machine learning model further comprises:

a third machine learning model configured to receive the text and the at least one initial image, the third machine learning model further configured to generate the at least one image based on the text and the at least one initial image, wherein the at least one image has a higher resolution than the at least one initial image.

13. The system of claim 9, wherein the text comprises input from the at least one user, and wherein the additional content comprises at least one portion of a live camera feed.

14. The system of claim 9, wherein the at least one image comprises different images generated based at least in part on a same text input by different users, and wherein the video comprises content overlaid on at least one subset of the plurality of different images.

15. A non-transitory computer-readable storage medium, storing computer-readable instructions that upon execution by a processor cause the processor to implement operations comprising:

receiving text, wherein the text is associated with a video to be created by at least one user;

receiving a style selection other than the text associated with the video to be created, wherein the style selection comprises input by selecting an icon corresponding to the desired style on an interface;

generating at least one image based at least in part on the text and the style selection using at least one machine learning model; and

generating the video based at least in part on the at least one image by overlay additional content onto the at least one image.

16. The non-transitory computer-readable storage medium of claim 15, wherein the style selection indicates at least one of a color, a texture, or an artistic style.

17. The non-transitory computer-readable storage medium of claim 15, wherein the at least one machine learning model comprises:

a first machine learning model configured to receive the text and generate at least one initial image based at least in part on the text; and

a second machine learning model configured to receive a plurality of images and generate a plurality of text captions corresponding to the plurality of images,

wherein the first machine learning model is trained on a dataset comprising the plurality of images and the plurality of corresponding text captions.

18. The non-transitory computer-readable storage medium of claim 17, wherein the at least one machine learning model comprises:

a third machine learning model configured to receive the text and the at least one initial image, the third machine learning model further configured to generate the at least one image based on the text and the at least one initial image, wherein the at least one image has a higher resolution than the at least one initial image.

19. The non-transitory computer-readable storage medium of claim 15, wherein the text comprises input from the at least one user, and wherein the additional content comprises at least one portion of a live camera feed.