CLAIMS:

1. A method comprising:

generating a plurality of sample images using a trained generative model;

for a sample image of the plurality of sample images, obtaining at least one attention map from the generative model, the at least one attention map being determined by the generative model for generating the sample image, an attention map indicating visual elements of an object within the sample image; and

performing training of a target model according to unsupervised learning at least based on the plurality of sample images and attention maps for the plurality of sample images, the target model being configured to perform an image processing task;

wherein the unsupervised learning comprises contrastive learning, and wherein performing training of the target model comprises: for a first sample image of the plurality of sample images that includes at least two objects,

for at least one feature of the first sample image by the target model, masking the at least one feature with at least two attention maps for the at least two objects, respectively, to obtain at least two masked features for the at least two objects, a masked feature comprising feature information related to an object;

constructing at least one positive sample pair and at least one negative sample pair from the at least two masked features, a positive sample pair comprising a pair of masked features for a same object, and a negative sample pair comprising a pair of masked features for a pair of different objects;

determining a contrastive loss for the first sample image based on at least one similarity between the at least one positive sample pair and at least one similarity between the at least one negative sample pair; and

training the target model based on the contrastive loss.

2. The method of claim 1, wherein generating the plurality of sample images comprises:

generating a plurality of sample images by providing a plurality of text prompts into a trained generative model, respectively, and

wherein an attention map for a sample image indicates visual elements of an object indicated by a text prompt.

3. The method of claim 2, further comprising:

generating the plurality of text prompts through at least one of the following:

filling at least one of a plurality of object class names into a text template, to obtain at least one text prompt, or

providing at least one of the plurality of object class names into a trained word-to-sentence model, to obtain at least one text prompt generated by the word-to-sentence model.

4. The method of claim 1, wherein masking the at least one feature with at least two attention maps for the at least two objects, respectively, to obtain at least two masked features for the at least two objects comprises:

applying a first cropping operation and a second cropping operation on the first sample image, respectively, to generate a first cropped image and a second cropped image;

applying the first cropping operation and the second cropping operation on at least two attention maps for the at least two objects, respectively, to obtain a first set of cropped attention maps for the first cropped image, and a second set of cropped attention maps for the second cropped image;

for a first feature of the first cropped image extracted by the target model, masking the first feature with the first set of cropped attention maps, respectively, to obtain a first set of masked features for the at least two objects; and

for a second feature of the second cropped image extracted by the target model, masking the second feature with the second set of cropped attention maps, respectively, to obtain a second set of masked features for the at least two objects.

5. The method of claim 1, wherein the unsupervised learning comprises masked modeling, and wherein performing training of the target model comprises: for a second sample image of the plurality of sample images,

masking at least one patch in the second sample image based on the at least one attention map for the second sample image; and

training the target model by performing masked modeling to reconstruct the at least one masked patch in the second sample image.

6. The method of claim 5, wherein an attention map for a sample image comprises importance scores of visual elements within the sample image with respect to an object; and wherein masking at least one patch in the second sample image comprises:

masking the at least one patch in the second sample image based on importance scores

comprised in the at least one attention map for the second sample image, the at least one masked patch corresponding to higher importance scores in an attention map than unmasked patches.

7. The method of claim 5, wherein masking at least one patch in the second sample image comprises:

in a first training iteration of the target model, masking a first ratio of patches among the plurality of patches in the second sample image, the first ratio being larger than a second ratio of patches masked in the second sample image in a training iteration earlier than the first training iteration.

8. The method of claim 1, wherein the unsupervised learning comprises vision-and-language pretraining, and wherein performing training of the target model comprises: for a third sample image of the plurality of sample images,

determining at least one location of at least object within the third sample image based on at least one attention map for the third sample image; and

performing the vision-and-language pretraining of the target model based at least in part on the third sample image and location information indicating the at least one location of the at least object within the third sample image.

9. The method of claim 8, wherein performing the vision-and-language pretraining of the target model comprises:

generating a text description to describe the at least one location of at least object within the third sample image; and

performing the vision-and-language pretraining of the target model based at least in part on the third sample image, the location information, and the text description.

10. The method of claim 1, wherein the generative model comprises a diffusion model for text-to-image generation.

11. A system, comprising:

at least one processor; and

at least one memory communicatively coupled to the at least one processor and comprising computer-readable instructions that upon execution by the at least one processor cause the at least one processor to perform acts comprising:

generating a plurality of sample images by providing a plurality of text prompts into a trained generative model, respectively;

for a sample image of the plurality of sample images, obtaining at least one attention map from the generative model, the at least one attention map being determined by the generative model for generating the sample image, an attention map indicating visual elements of an object within the sample image; and

performing training of a target model according to unsupervised learning at least based on the plurality of sample images and attention maps for the plurality of sample images, the target model being configured to perform an image processing task;

wherein the unsupervised learning comprises contrastive learning, and wherein performing training of the target model comprises: for a first sample image of the plurality of sample images that includes at least two objects,

for at least one feature of the first sample image by the target model, masking the at least one feature with at least two attention maps for the at least two objects, respectively, to obtain at least two masked features for the at least two objects, a masked feature comprising feature information related to an object;

constructing at least one positive sample pair and at least one negative sample pair from the at least two masked features, a positive sample pair comprising a pair of masked features for a same object, and a negative sample pair comprising a pair of masked features for a pair of different objects;

determining a contrastive loss for the first sample image based on at least one similarity between the at least one positive sample pair and at least one similarity between the at least one negative sample pair; and

training the target model based on the contrastive loss.

12. The system of claim 11, wherein generating the plurality of sample images comprises:

generating a plurality of sample images by providing a plurality of text prompts into a trained generative model, respectively, and

wherein an attention map for a sample image indicates visual elements of an object indicated by a text prompt.

13. The system of claim 11, wherein masking the at least one feature with at least two attention maps for the at least two objects, respectively, to obtain at least two masked features

for the at least two objects comprises:

applying a first cropping operation and a second cropping operation on the first sample image, respectively, to generate a first cropped image and a second cropped image;

applying the first cropping operation and the second cropping operation on at least two attention maps for the at least two objects, respectively, to obtain a first set of cropped attention maps for the first cropped image, and a second set of cropped attention maps for the second cropped image;

for a first feature of the first cropped image extracted by the target model, masking the first feature with the first set of cropped attention maps, respectively, to obtain a first set of masked features for the at least two objects; and

for a second feature of the second cropped image extracted by the target model, masking the second feature with the second set of cropped attention maps, respectively, to obtain a second set of masked features for the at least two objects.

14. The system of claim 11, wherein the unsupervised learning comprises masked modeling, and wherein performing training of the target model comprises: for a second sample image of the plurality of sample images,

masking at least one patch in the second sample image based on the at least one attention map for the second sample image; and

training the target model by performing masked modeling to reconstruct the at least one masked patch in the second sample image.

15. The system of claim 14, wherein an attention map for a sample image comprises importance scores of visual elements within the sample image with respect to an object; and wherein masking at least one patch in the second sample image comprises:

masking the at least one patch in the second sample image based on importance scores comprised in the at least one attention map for the second sample image, the at least one masked patch corresponding to higher importance scores in an attention map than unmasked patches.

16. The system of claim 14, wherein masking at least one patch in the second sample image comprises:

in a first training iteration of the target model, masking a first ratio of patches among the plurality of patches in the second sample image, the first ratio being larger than a second ratio of patches masked in the second sample image in a training iteration earlier than the first

training iteration.

17. The system of claim 11, wherein the unsupervised learning comprises vision-and-language pretraining, and wherein performing training of the target model comprises: for a third sample image of the plurality of sample images,

determining at least one location of at least object within the third sample image based on at least one attention map for the third sample image; and

performing the vision-and-language pretraining of the target model based at least in part on the third sample image and location information indicating the at least one location of the at least object within the third sample image.

18. A computer-readable storage medium, storing computer-readable instructions that upon execution by a device cause the device to perform acts comprising:

generating a plurality of sample images by providing a plurality of text prompts into a trained generative model, respectively;

for a sample image of the plurality of sample images, obtaining at least one attention map from the generative model, the at least one attention map being determined by the generative model for generating the sample image, an attention map indicating visual elements of an object within the sample image; and

performing training of a target model according to unsupervised learning at least based on the plurality of sample images and attention maps for the plurality of sample images, the target model being configured to perform an image processing task;

wherein the unsupervised learning comprises contrastive learning, and wherein performing training of the target model comprises: for a first sample image of the plurality of sample images that includes at least two objects,

for at least one feature of the first sample image by the target model, masking the at least one feature with at least two attention maps for the at least two objects, respectively, to obtain at least two masked features for the at least two objects, a masked feature comprising feature information related to an object;

constructing at least one positive sample pair and at least one negative sample pair from the at least two masked features, a positive sample pair comprising a pair of masked features for a same object, and a negative sample pair comprising a pair of masked features for a pair of different objects;

determining a contrastive loss for the first sample image based on at least one similarity

between the at least one positive sample pair and at least one similarity between the at least one negative sample pair; and

training the target model based on the contrastive loss.