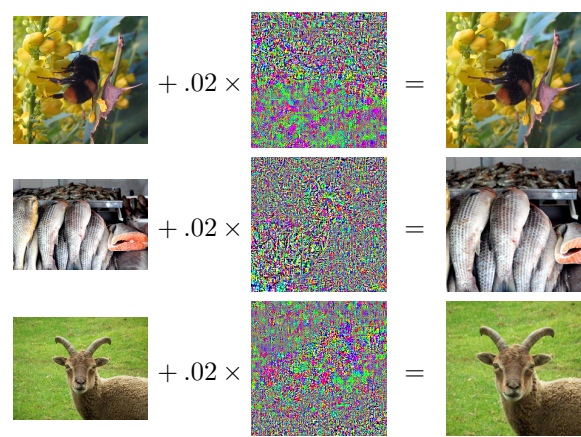


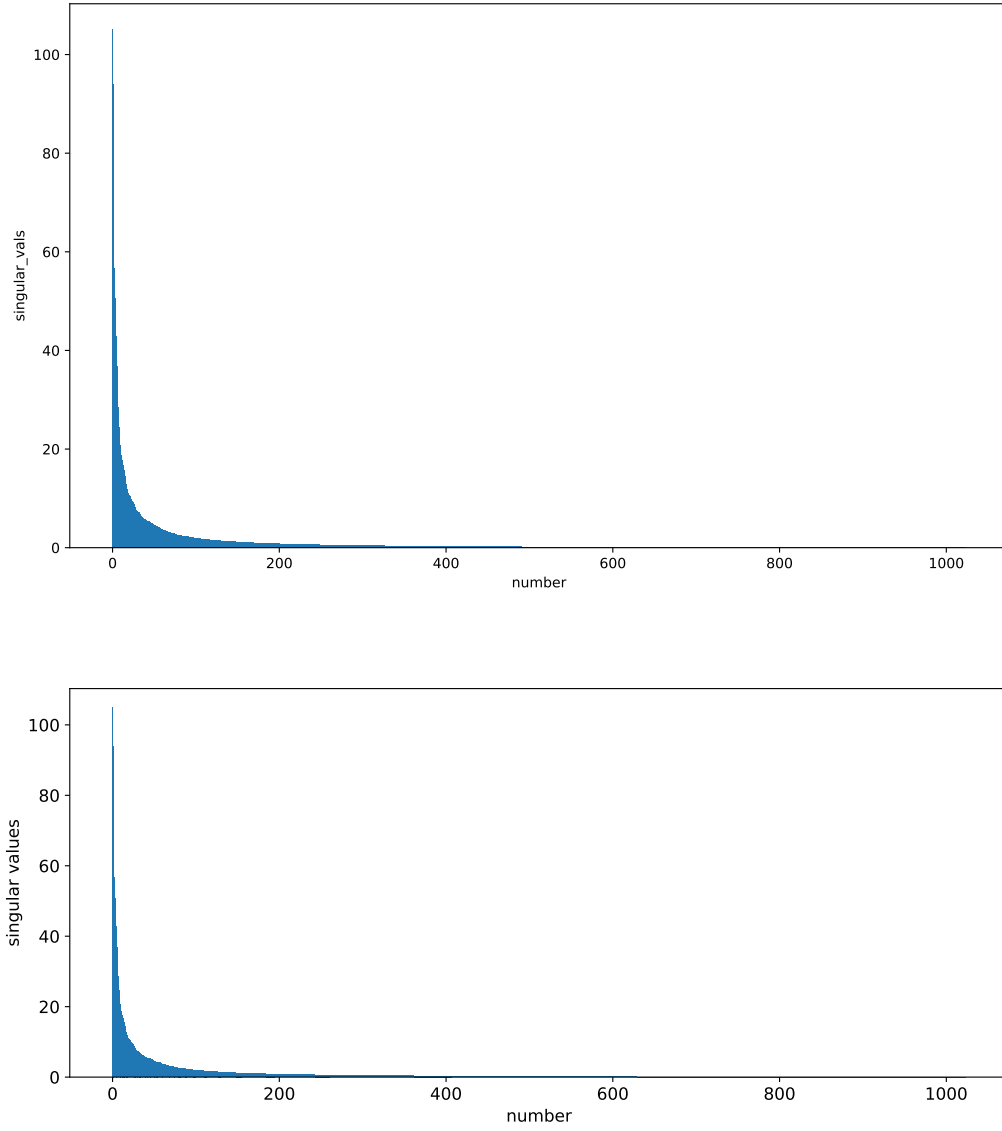
Supplementary Material

A. ADDITIONAL RESULTS

Here we provide more details and additional information about the results we have included in the main text.



**Fig. 5.** Pixel differences between the two images in each of the three pairs in Fig. 1; they are multiplied by 50 for visualization.



**Fig. 6.** Local structures of the embedding space. (top) The singular values of the Jacobian Matrix for Fig. 1(a); (bottom) The singular values of the Jacobian Matrix for Fig. 1(b).



**Fig. 7.** Interpolated Images for Fig. 4.

Fig. 6 shows the singular values of the Jacobian matrix for the original (Fig. 1(a)) and modified image (Fig. 1(e)) pair. Fig. 7 shows the interpolated images along the path.

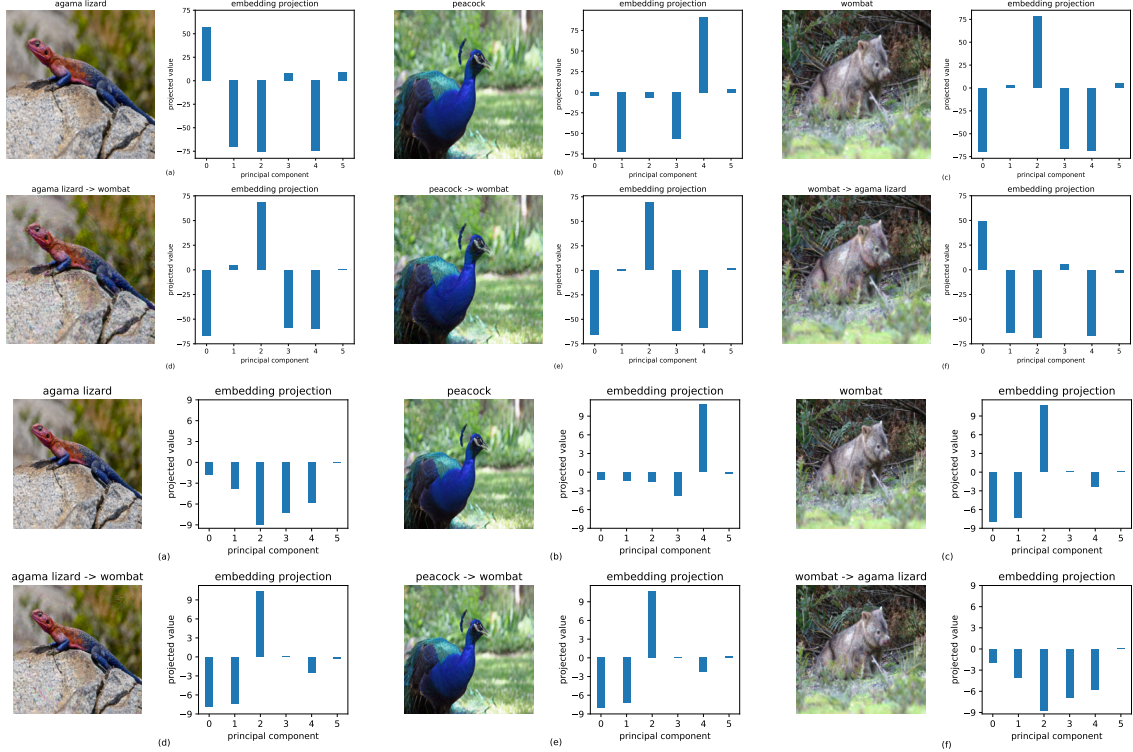
To further demonstrate the effectiveness of the gradient procedure to match embeddings, we have applied the procedure to numerous images from different sources. As random images are typical in the input image space, we have applied the procedure

to match a specified embedding from randomly generated images. Fig. 9 shows that we can match the embeddings of images from a random image; These results, along with outcomes from other datasets, demonstrate the efficacy of our technique across all the images we have utilized.

In the main paper, the results are generated using the pre-trained ImageBind [12] model, which utilizes a pre-trained CLIP model (ViT-H-14). As the framework does not rely on the specifics of the ImageBind, it is effective for other models and datasets as well. To demonstrate that our framework works equally well with other variants, Fig. 10 shows the results on several different variants of the original vision transformer models<sup>8</sup>. To further showcase the model-agnostic nature of our techniques, we conduct experiments with diverse vision transformer models, including DeiT, ViTMAE, and ViTMSN. Please refer to Fig. 8 and Fig. 11 for detailed results.

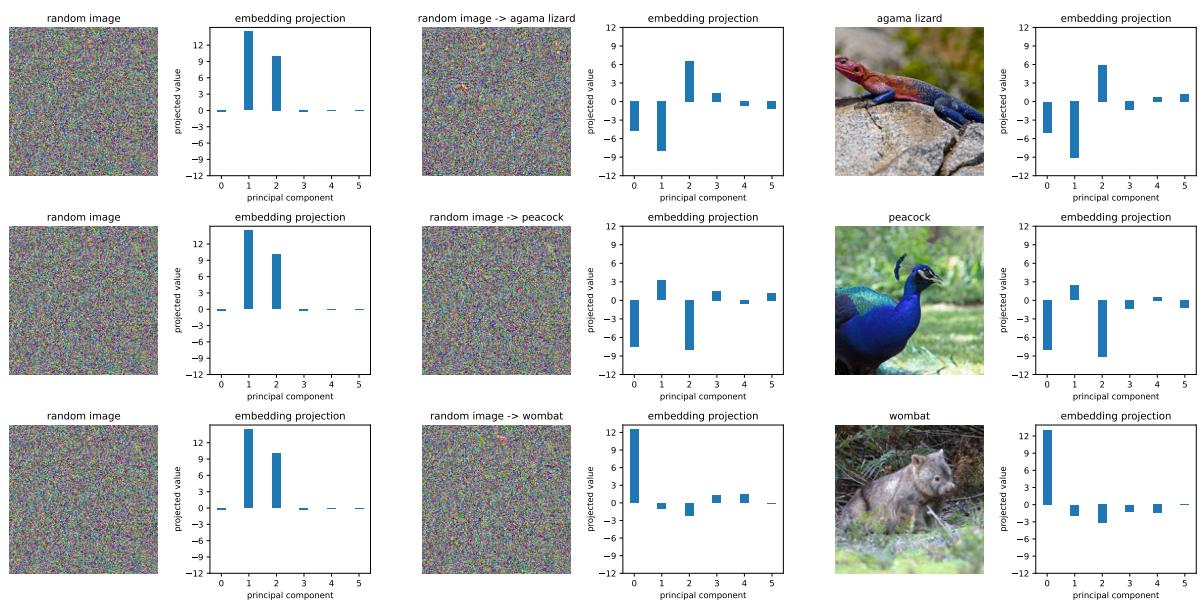
Fig. 14 provides more examples on Google Open Images, where visually indistinguishable images have very different embeddings and consequently are classified into other classes. In contrast, visually very different images have very similar embeddings, aligned to the embedding of a particular image and classified into the corresponding class. Additionally, in Fig. 13 and Fig. 12, we present further examples applying our proposed framework to the MS-COCO and ImageNet datasets, affirming the dataset-agnostic nature of our approach.

Fig. 15 provides the original images from ImageNet used in Fig. 12. Similarly, Fig. 16 shows the original images from MS-COCO and Open Images dataset used to generate the Fig. 13 and Fig. 14.



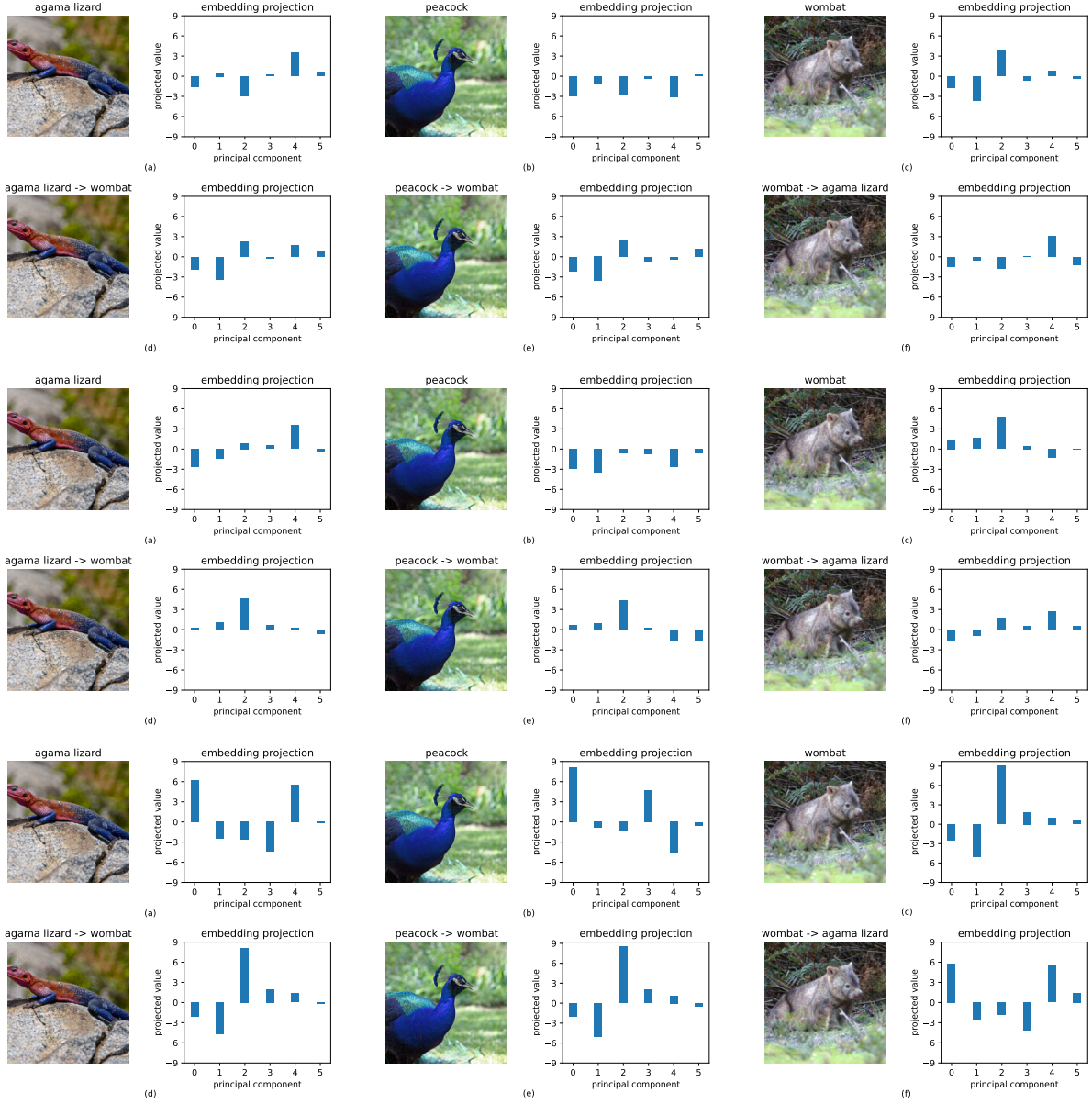
**Fig. 8.** Examples obtained while the proposed framework is applied on different vision transformer models, such as (top two rows) BEiT, and (the next two rows) Swin Transformer. The results are given in the same format as depicted in Fig. 1. Additional plots for other models are also consistent and added to the Supplemental Material. The example demonstrates that the method is model-agnostic.

<sup>8</sup><https://github.com/openai/CLIP>

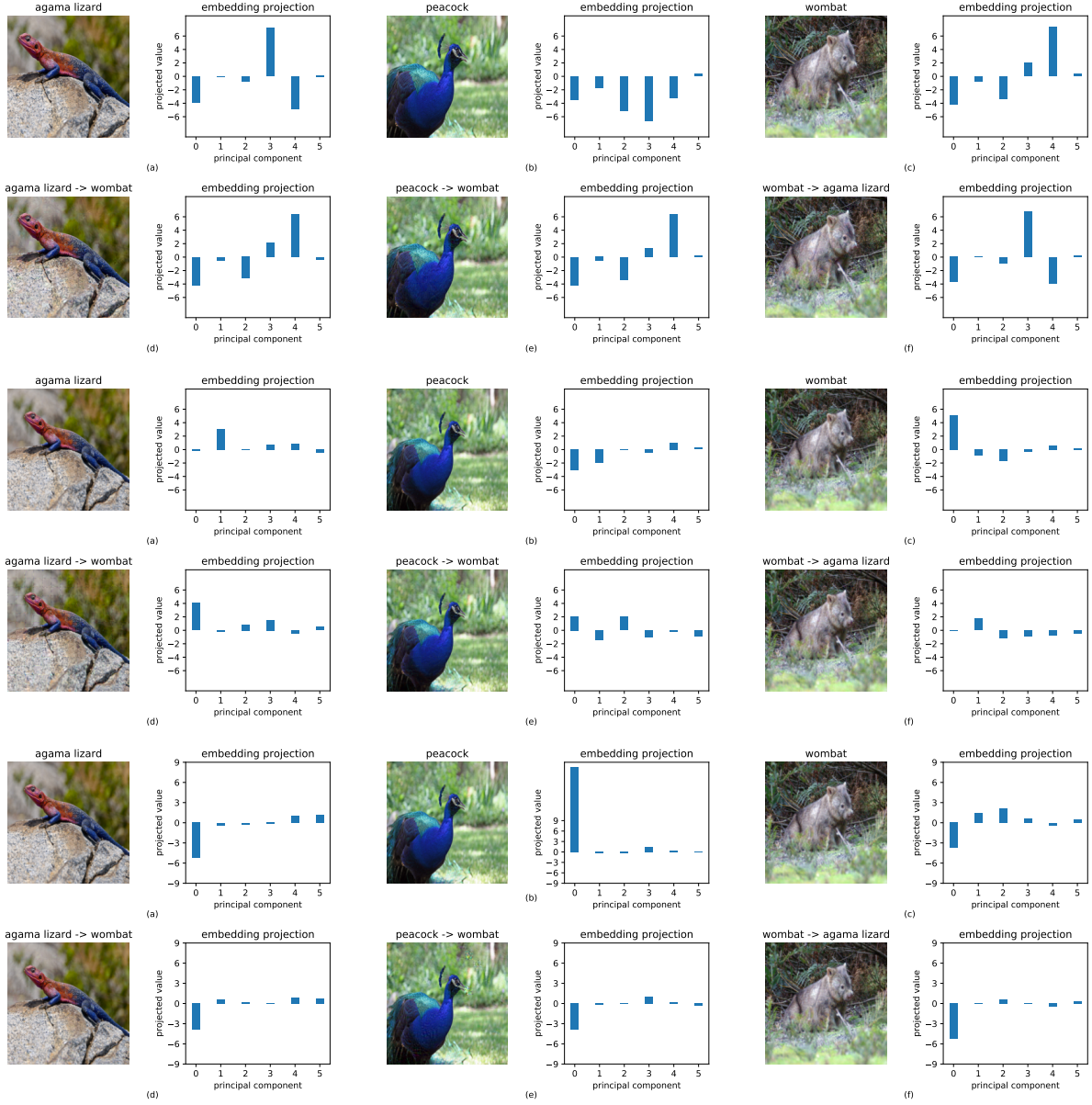


**Fig. 9.** Example of random image (left) that matches a target embedding (right), with the final image shown in the middle.

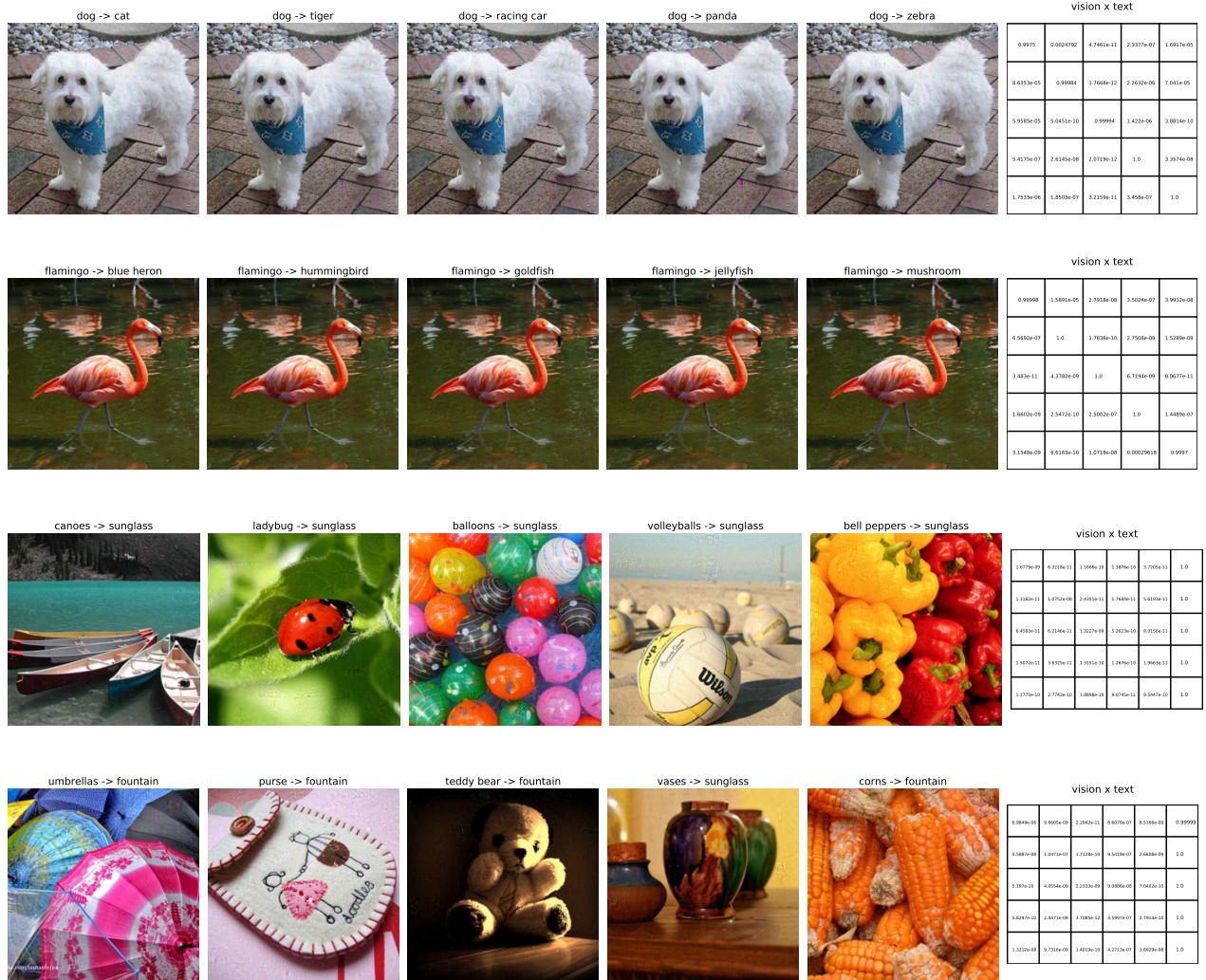




**Fig. 10.** More examples from ImageNet obtained using the proposed framework with different variants of the original vision transformer, such as (top) ViT-B-16, which has the embedding dimension of 512, (center) ViT-B-32, which has the embedding dimension of 512, (bottom) ViT-L-14 which has the embedding dimension of 768.

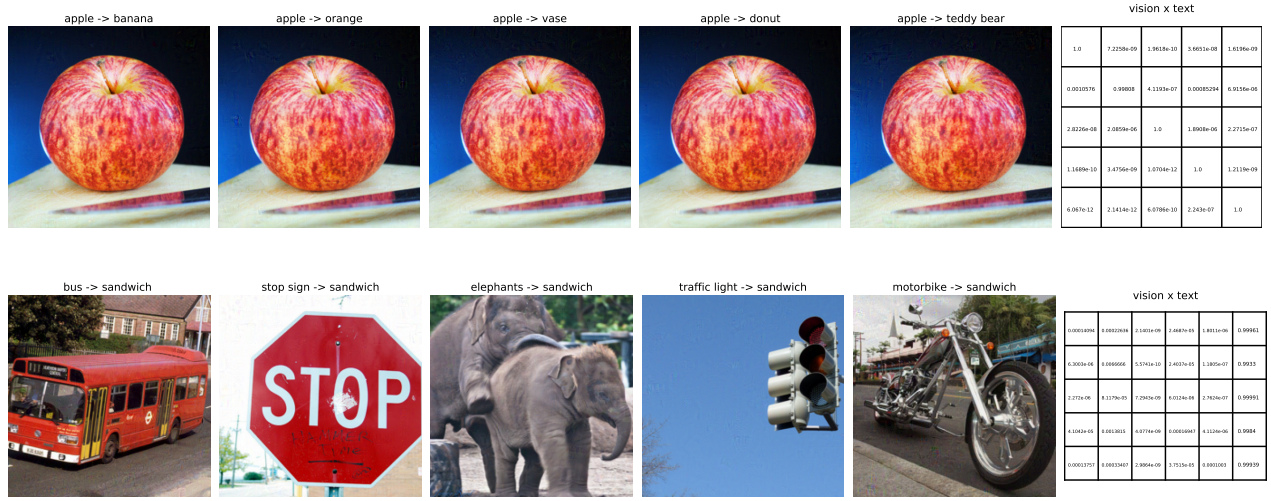


**Fig. 11.** Same as Fig. 1 and Fig. 8, in support of demonstrating that the proposed framework is model-agnostic; shown for different other vision transformer models, such as (top two rows) DeiT, (middle two rows) ViTMAE and (the next two rows) ViTMSN.

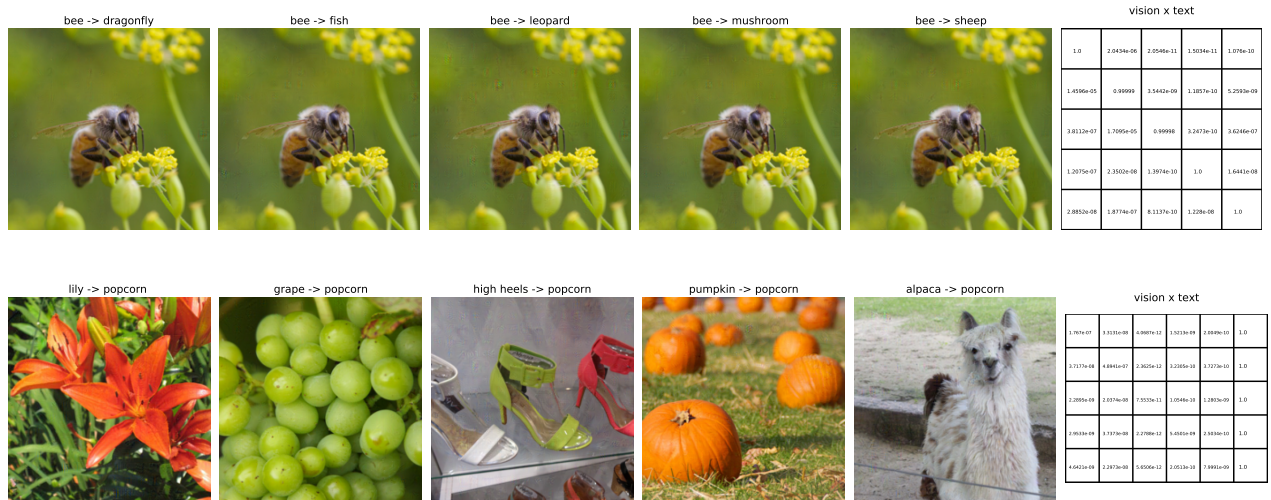


**Fig. 12.** (first row) Additional examples where visually indistinguishable images have very different embeddings and consequently are classified to other classes as in Fig. 1. Dog images are classified as a cat, a tiger, a racing car, a panda, and a zebra. (second row) Similar as first row, flamingo images are classified as a heron, hummingbird, goldfish, jellyfish, and mushroom. (third row) Visually very different images (e.g., some canoes, a ladybug, some balloons, some volleyballs, some bell peppers) have very similar embeddings and are classified as sunglass. (fourth row) Similar as third row, different images (e.g., some umbrellas, a purse, a teddy bear, some vases, some corns) are classified as fountain. The examples are strictly randomly chosen. There is no postselection involved.





**Fig. 13.** More examples involving MS-COCO dataset. (top) Visually indistinguishable images have very different representations via embedding alignment with the corresponding images and therefore very different classification outcomes. (bottom) Visually very different images have very similar embeddings, aligned to the embedding of a specific image and classified into the corresponding class. Again the samples are randomly chosen.

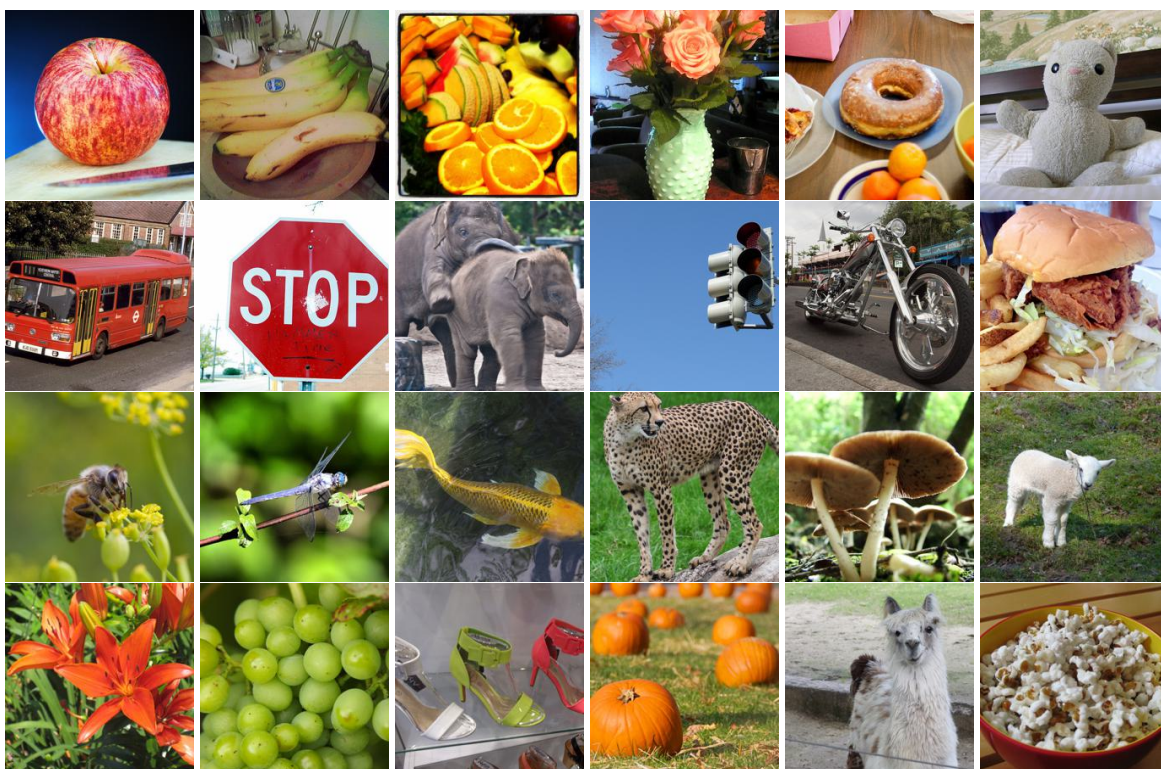


**Fig. 14.** More examples involving Open Images dataset having high-resolution images. (top) Visually indistinguishable images have very different representations via embedding alignment with the corresponding images and therefore very different classification outcomes. (bottom) Visually very different images have very similar embeddings, aligned to the embedding of a specific image and classified into the corresponding class. The samples are randomly chosen.





**Fig. 15.** The original images from ImageNet corresponds to Fig 12.



**Fig. 16.** The original images correspond to Fig. 13 and Fig 14. (first two rows) MS-COCO, and (the next two rows) Google Open Images.