SUPPLEMENTARY MATERIAL FOR SUBMISSION:

RECONSTRUCTING TRAINING DATA FROM REAL WORLD MODELS TRAINED WITH TRANSFER LEARNING

Anonymous authors
Paper under double-blind review

1 FULL RESULTS FOR FIG.5B FROM MAIN PAPER

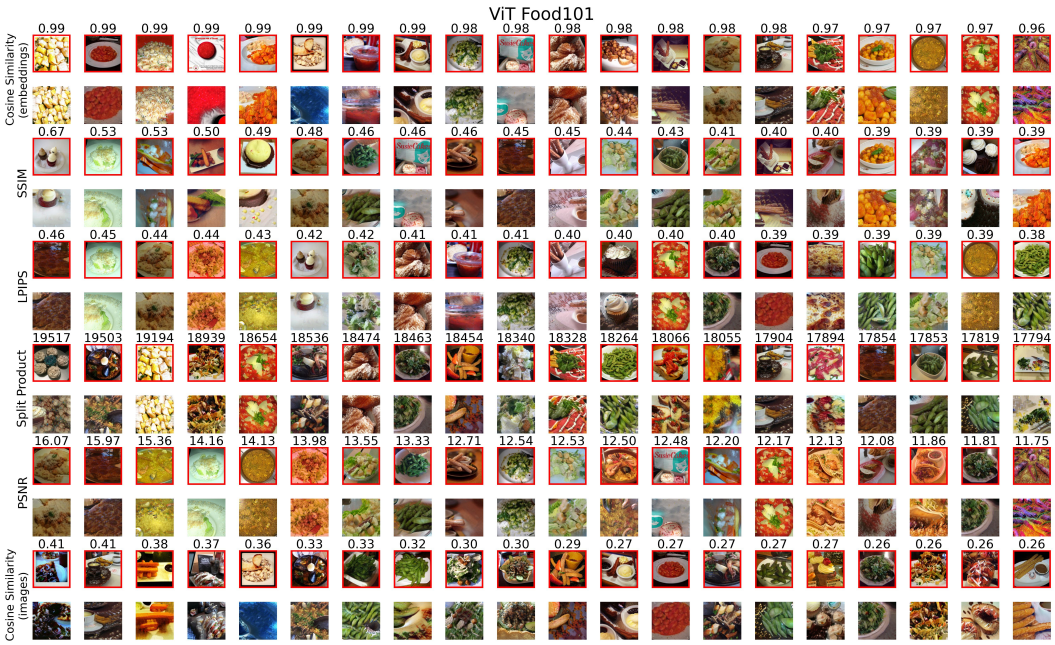
In Figure 5b in the main paper we show reconstructed samples, sorted according to their reconstruction-quality as measured by cosine-similarity between their image embeddings. The results shown there are obtained from a model trained on DINO embeddings on Food101 dataset.

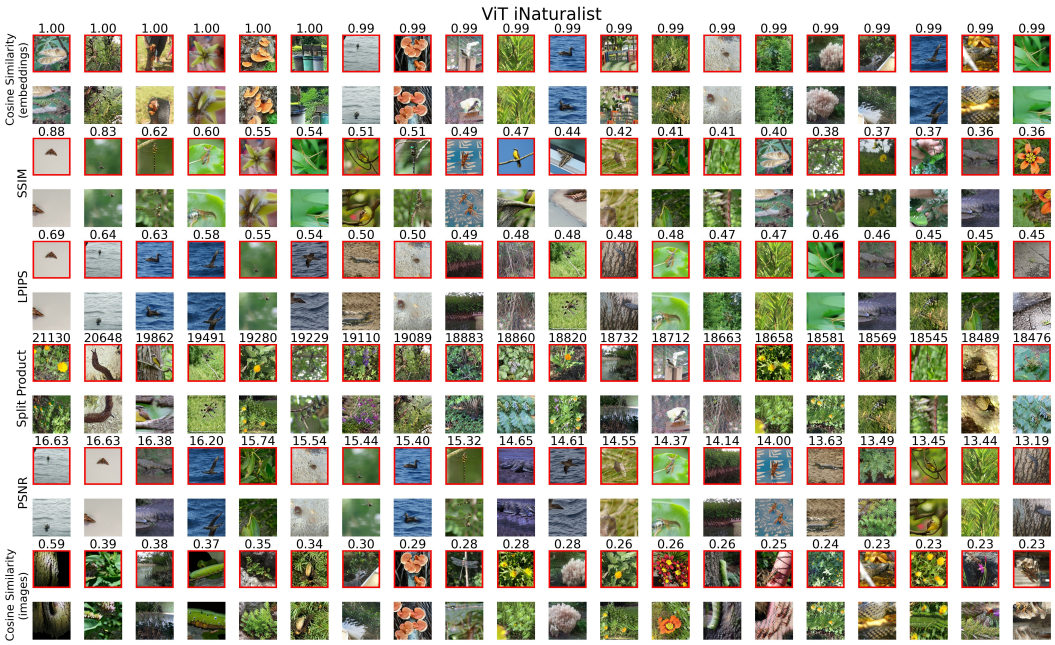
Below we provide the complete results for this type of evaluation, namely, for each model from Figure 3, and for each reconstruction-quality metric, we show the "best" reconstructed samples according to this metric (by sorting them).

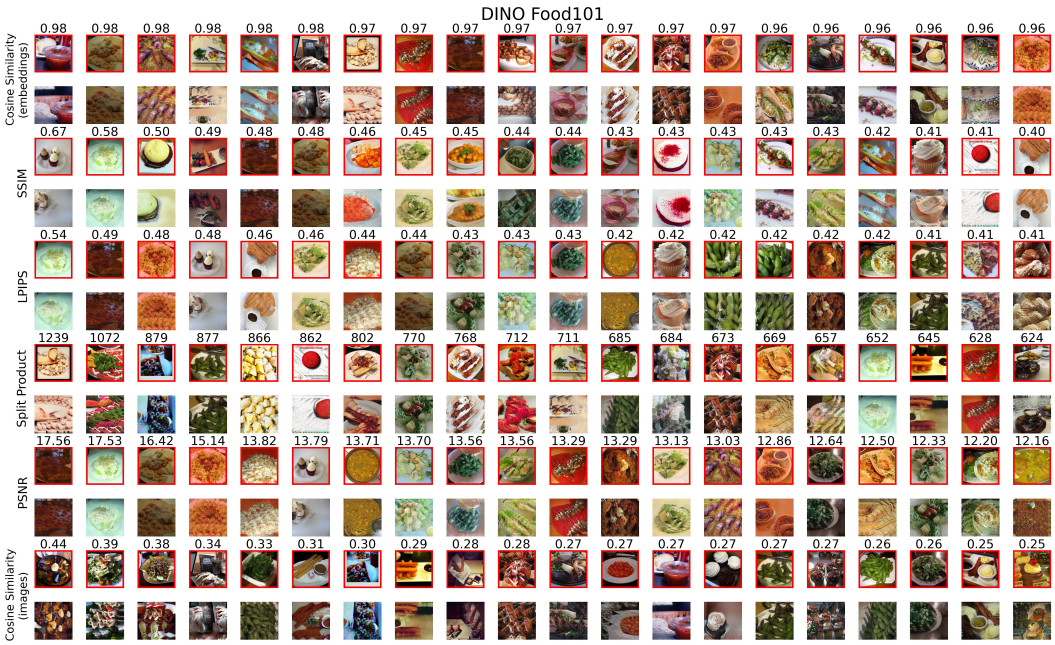
In total there are 8 models: trained on 4 backbones (ViT, DINO, DINOv2 and CLIP) and on 2 datasets (Food101 and iNaturalist), as described in details in the Results Section in the main paper. And for each model we show the sorted results for a total of 6 choices for reconstruction-quality: Cosine-Similarity in Embedding space plus 5 metrics in Image space: SSIM, LPIPS, Split-Product, PSNR and Cosine-Similarity (Image space).

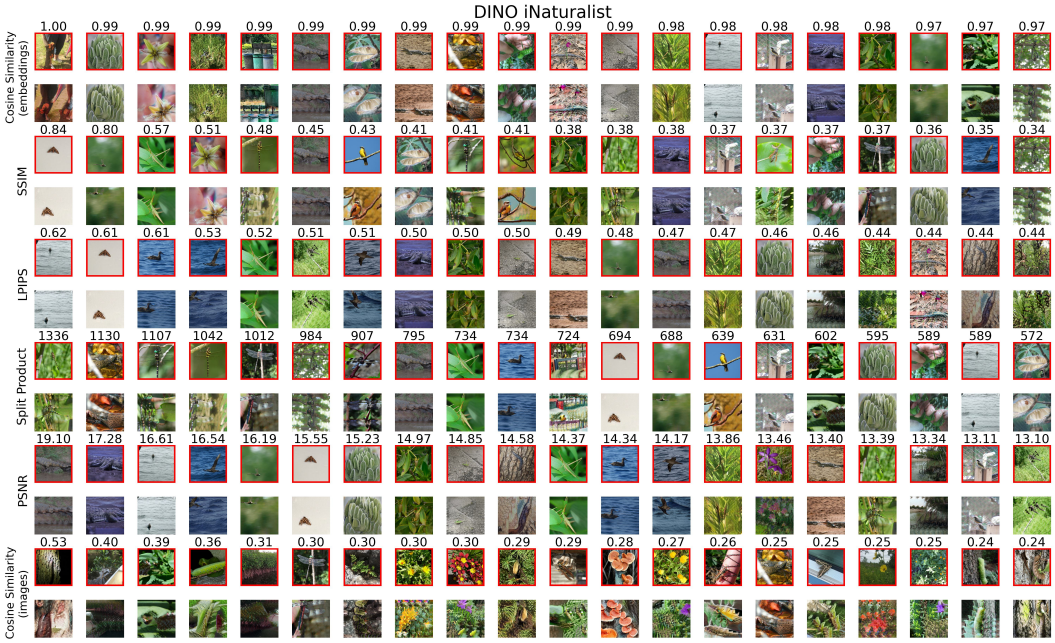
In each Figure below, images with RED borderline are original training images, and the image below them is their nearest reconstructed image (as measured by the cosine-similarity between their embeddings). Note: in all cases, the matching between a training image and its reconstructed image is the same. the only difference is between the way they are sorted – which is done using the metric (as written in the left side of each row).

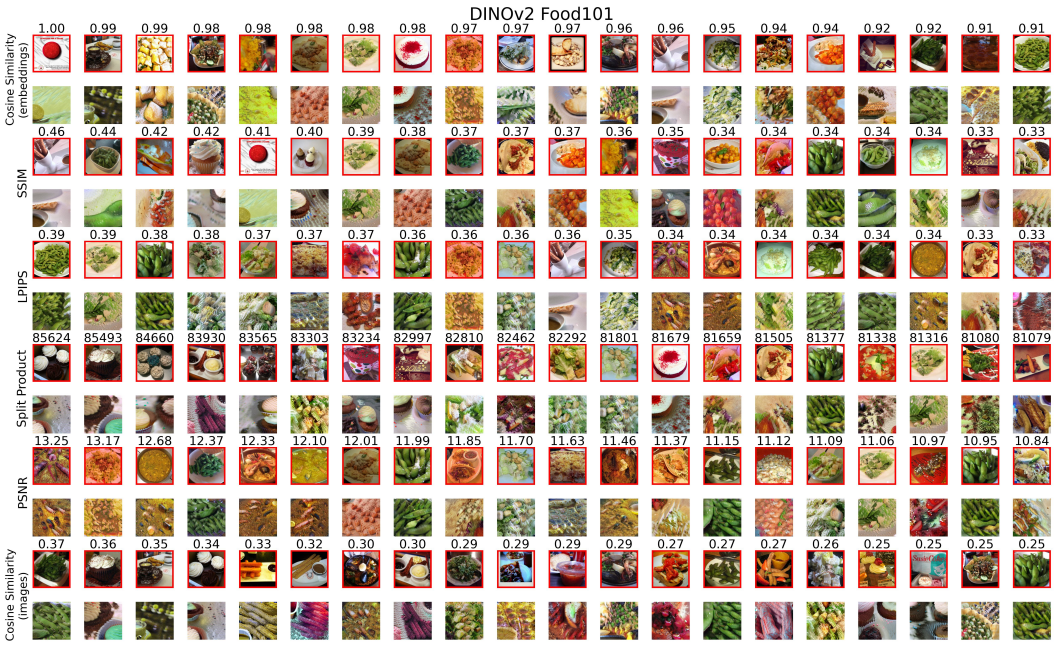
We are aware that there may be some sampling bias in these results. However, since in our work, the inversion part is time-intensive, we must choose which embeddings to invert, where in our work we use the cosine-similarity for that, as detailed in the paper. It is not feasible to use an image-metric for this goal, because this would mean first inverting all candidate embeddings (usually 25k-50k of them), which is not feasible.

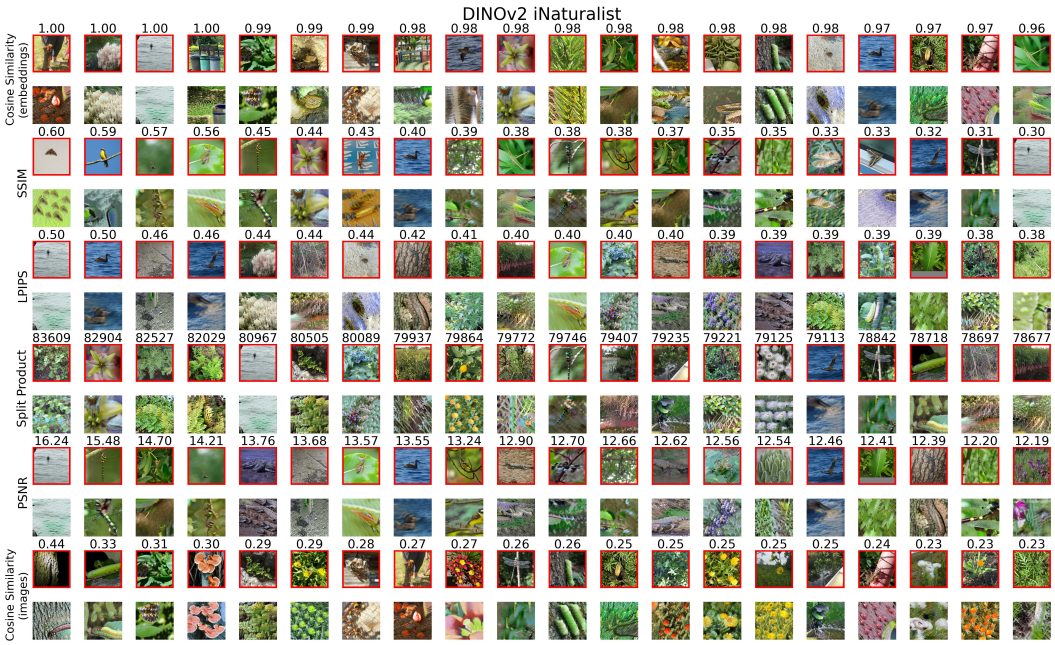

ViT Food101
Cosine Similarity (embeddings)
SSIM
LPIPS
Split Product
PSNR
Cosine Similarity (images)


ViT iNaturalist
Cosine Similarity (embeddings)
SSIM
LPIPS
Split Product
PSNR
Cosine Similarity (images)

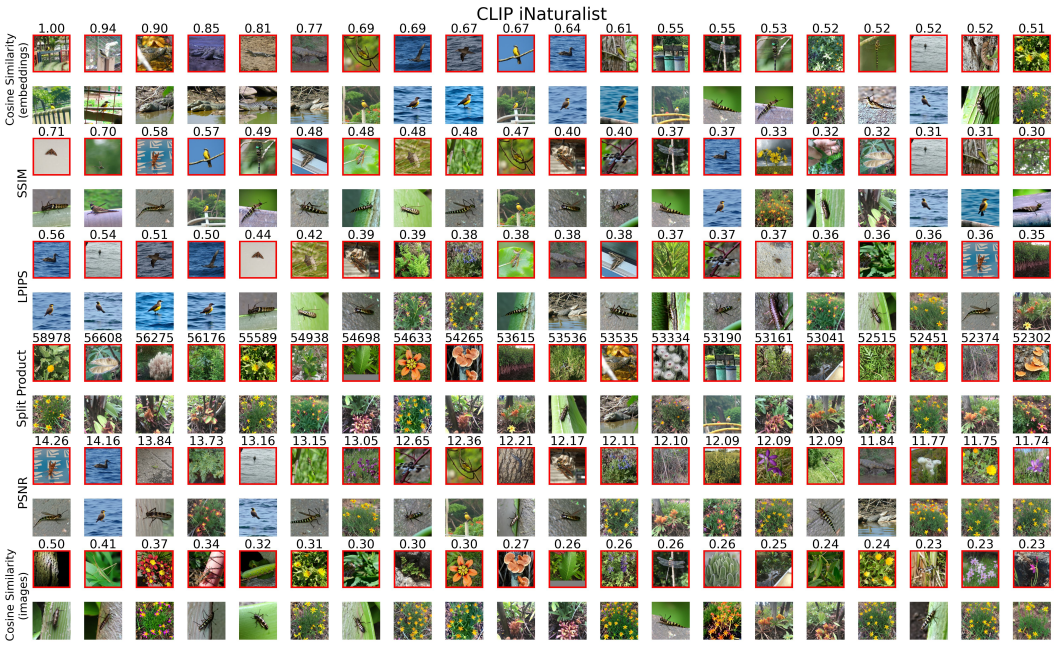

DINO Food101
Cosine Similarity (embeddings)
SSIM
LPIPS
Split Product
PSNR
Cosine Similarity (images)


DINO iNaturalist
Cosine Similarity (embeddings)
SSIM
LPIPS
Split Product
PSNR
Cosine Similarity (images)


DINOv2 Food101
Cosine Similarity (embeddings)
SSIM
LPIPS
Split Product
PSNR
Cosine Similarity (images)


DINOv2 iNaturalist
Cosine Similarity (embeddings)
SSIM
LPIPS
Split Product
PSNR
Cosine Similarity (images)


CLIP Food101
Cosine Similarity (embeddings)
1.00 1.00 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.98 0.98 0.98 0.98 0.98 0.97 0.97 0.97
SSIM
0.52 0.48 0.48 0.48 0.46 0.45 0.44 0.44 0.43 0.43 0.42 0.42 0.42 0.42 0.41 0.40 0.40 0.40 0.40 0.39
LPIPS
0.45 0.43 0.43 0.41 0.41 0.41 0.40 0.40 0.39 0.39 0.38 0.38 0.38 0.38 0.37 0.37 0.37 0.37 0.37 0.36
Split Product
56432 56183 56130 56019 55962 55651 55229 55102 55054 54978 54936 54703 54573 54559 54469 54427 54364 54160 54044 53957
PSNR
12.74 12.59 12.53 11.70 11.54 11.45 11.39 10.93 10.78 10.75 10.74 10.70 10.59 10.43 10.37 10.35 10.35 10.35 10.28 10.28
Cosine Similarity (images)
0.50 0.42 0.41 0.38 0.36 0.36 0.35 0.34 0.33 0.33 0.33 0.33 0.32 0.32 0.31 0.31 0.31 0.30 0.30 0.30


CLIP iNaturalist
Cosine Similarity (embeddings)
1.00 0.94 0.90 0.85 0.81 0.77 0.69 0.69 0.67 0.67 0.64 0.61 0.55 0.55 0.53 0.52 0.52 0.52 0.52 0.51
SSIM
0.71 0.70 0.58 0.57 0.49 0.48 0.48 0.48 0.48 0.47 0.40 0.40 0.37 0.37 0.33 0.32 0.32 0.31 0.31 0.30
LPIPS
0.56 0.54 0.51 0.50 0.44 0.42 0.39 0.39 0.38 0.38 0.38 0.38 0.37 0.37 0.37 0.36 0.36 0.36 0.36 0.35
Split Product
58978 56608 56275 56176 55589 54938 54698 54633 54265 53615 53536 53535 53334 53190 53161 53041 52515 52451 52374 52302
PSNR
14.26 14.16 13.84 13.73 13.16 13.15 13.05 12.65 12.36 12.21 12.17 12.11 12.10 12.09 12.09 12.09 11.84 11.77 11.75 11.74
Cosine Similarity (images)
0.50 0.41 0.37 0.34 0.32 0.31 0.30 0.30 0.30 0.27 0.26 0.26 0.26 0.26 0.25 0.24 0.24 0.23 0.23 0.23