Reconstructing Animals *and* the Wild

Supplementary Material

A. Training Details

We finetune the LLaMA 1-based Vicuna 1.3 model with LoRA [1]. We use the HuggingFace Transformers and PEFT libraries, along with DeepSpeed `ZeRO-2` [5]. In all experiments, we use a `lora_r` of `128`, a `lora_alpha` of `256`, a LoRA learning rate of `2e-05`, an input linear projector learning rate of `2e-05`, a learning rate of `0.0002` for the CLIP and numeric heads, and a cosine learning-rate schedule. All models are trained with a batch size of `32` for `100,000` steps with `bfloat16` mixed-precision training. We use the CLIP encoder from [1], the DINOv2 encoder from [2], and the BioCLIP encoder from [3]. The weight of the rotation-matrix MSE is `1`, the cosine-similarity loss applied on the embedding `10`, and the embedding norm `0.001`.

B. Object-Level Evaluation

We additionally perform an object-level evaluation on the synthetic in-distribution testing data. To do so, we match ground-truth objects to predictions by the nearest distance between them. Results are reported in Tab. S.1. We observe that the text-based model, while performing worst of the models in estimating assets of the correct type/category, scores highest on layout metrics. We suggest that the additional conditioning the model has during training (i.e., the multiple asset name tokens) reduces train-time ambiguity, allowing it to learn to estimate with greater positional precision, at the cost of semantic similarity. However, we note that while shape-estimation performance appears comparable in-distribution between the CLIP-based model and the IG-LLM baseline, the accuracy is skewed: evaluating at five objects the balanced accuracy of IG-LLM is 52% while that of the CLIP variant is 80%.

While our investigation was focused on holistic compositional scene reconstruction, for which to our knowledge only a true comparison against IG-LLM can be made in a straightforward way, we opt to also evaluate against YOLOX-6D-Pose [4], a recent state-of-the-art single-shot 6D-pose-estimation method, on the object-based portion of our task. We train the extra-large version of YOLOX-6D-Pose for 30 epochs using all recommended data augmentations and the RAW train set, and evaluate the model on the same test set as used above. However, we observe that it fails to capture the task, identifying objects, but with appearances that are not semantically appropriate. We record results in Tab. S.1, and find that it falls excessively behind the other methods. We suggest that the issue observed is similar in nature to that of the training of the text-based model, but that the model is worse-off because there is no sequence to be conditioned on during training.

[1] https://huggingface.co/laion/CLIP-convnext_xxlarge-laion2B-s34B-b82K-augreg-soup

[2] https://huggingface.co/facebook/dinov2-giant

[3] https://huggingface.co/imageomics/bioclip

Figure S.1. **Crop Retrieval.** Nearest assets retrieved by CLIP embeddings of crops. The most-salient objects were hand cropped.

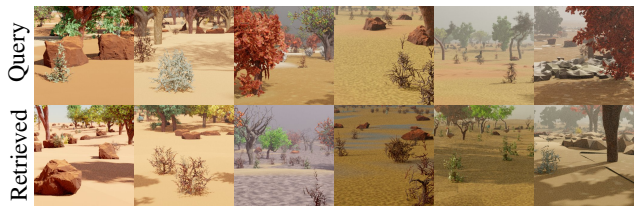

Figure S.2. **Scene Retrieval.** Training images were retrieved by closest scene-level CLIP similarity for queried testing images.

C. CLEVR-CoGenT

We evaluate the ability of the CLIP-augmented model to generalize compositionally. To do so, we employ the CLEVR [2] dataset. CLEVR is a procedurally generated benchmark of renderings of primitive objects with various discrete attribute combinations. It contains a subset, known as CLEVR-CoGenT, where all cubes in training images are gray, blue, brown, or yellow, and all cylinders red, green, purple, or cyan. During testing, these combinations are flipped. See IG-LLM [3] for additional details on the base setup.

We train our model on a simplified format, where the textual `color` and `shape` attributes are replaced by a single `[APP]` token. The model is supervised via the same losses as described in the methods section, but without a loss on rotations, which are not applicable to the evaluation. We train on four-thousand images and forty-thousand steps, matching the evaluation as performed in IG-LLM against NS-VQA [6], but decrease the effective batch size to eight.

The CLIP-augmented model achieves greater shape-estimation performance than NS-VQA in the OOD setting,

	↓Pos. L2	↓Geod.	↓Ht. MSE	↑Type Acc.	↓Pixel MAE	↑CLIP Sim.
5 Obj.						
CLIP	3.150	0.072	8.243	0.844	1759.847	**0.873**
Text	**2.810**	**0.067**	**6.883**	0.836	**1551.264**	0.862
BioCLIP	3.518	0.070	8.022	**0.856**	1934.479	0.852
DINOv2	3.130	**0.067**	7.791	**0.856**	1815.325	0.867
YOLOX-6D-Pose	5.472	1.270	22.185	0.519	N/A	0.764
10 Obj.						
CLIP	4.285	0.098	7.825	0.839	1195.519	**0.858**
Text	**3.703**	**0.089**	**6.640**	0.825	**1062.849**	0.846
BioCLIP	4.685	0.096	7.772	**0.847**	1320.637	0.838
DINOv2	4.183	0.098	7.567	0.846	1248.744	0.851
YOLOX-6D-Pose	5.139	1.232	19.282	0.566	N/A	0.763
15 Obj.						
CLIP	5.143	0.117	8.150	0.823	985.282	**0.846**
Text	**4.429**	**0.110**	**7.143**	0.808	**881.454**	0.833
BioCLIP	5.640	0.117	8.629	0.817	1116.113	0.824
DINOv2	5.016	0.117	8.118	**0.825**	1025.072	0.837
YOLOX-6D-Pose	5.079	1.205	17.377	0.589	N/A	0.762

Table S.1. **Object-Level Evaluation.** We additionally report quantitative in-distribution results at an object level. To compute each metric, ground-truth objects are matched to the nearest prediction. Pos. L2 represents Euclidean distance, Geod. stands for geodesic distance applied on the estimated rotations, Ht. MSE is measured on estimated object height, Pixel MAE is the mean-absolute error between predictions of pixel count, and CLIP Sim. is the asset-wise distance computed on embeddings of asset images. Results are recorded for the first five, first ten, and first fifteen ground-truth objects. CLIP similarity is shown separately because it is included in the training objective of the CLIP-based model.

	ID			OOD		
	CLIP	IG-LLM	NS-VQA	CLIP	IG-LLM	NS-VQA
↓L2	0.19	0.21	0.18	0.20	0.17	0.18
↑Size	99.63	99.71	100.00	99.53	99.80	100.00
↑Color	87.00	99.58	100.00	83.25	98.14	99.95
↑Shape	99.25	99.51	100.00	43.15	93.14	33.88

Table S.2. **CLEVR-CoGenT Results.**

but that it lags behind IG-LLM (Tab. S.2). We hypothesize that the CLIP-projection head is over-parametrized relative to the number of unique embeddings. We compute the average cosine similarity between embeddings produced from CLEVR shapes as 0.81, which contrasts to the mean similarity value of 0.59 of RAW assets. See a computed similarity matrix in Fig. S.3.

D. Data-Efficiency Evaluation

In Tab. S.3 we evaluate the data efficiency of the framework, training on 10,000, 100,000, and 1,000,000 samples. We notice consistent performance gains as the amount of data increases, indicating that model performance could likely be increased were the amount of training data further increased.

	↓LPIPS	↑ S_{CLIP}	↑ S_{BioCLIP}	↑ S_{DINOv2}
10k	0.829	0.698	0.327	0.786
100k	0.697	0.794	0.482	0.834
1M	**0.598**	**0.815**	**0.539**	**0.858**

Table S.3. **Data Efficiency.** We observe a consistent increase in model performance as the number of training samples is increased, indicating that the benefit of adding additional training data may not have saturated.

E. Asset Separability

In Fig. S.4 we show retrieval samples using ground-truth embeddings for each of the visual encoders. The 'input' comes from a held-out query set and the retrieved assets

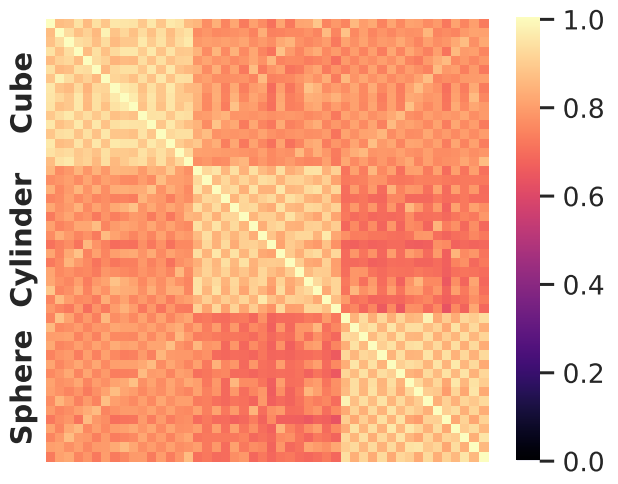

Figure S.3. **CLEVR-CoGenT Embedding Similarity.**

are from the set of assets used in the training samples. The retrievals in the figure represent an upper bound on retrieval performance as the model is trained with these embeddings as targets.

In Fig. S.5 we show a scatter plot of the first two PCA components of the CLIP embeddings of a set of trees. The relative grouping of the projections suggests that CLIP embeddings can be used to distinguish between asset instances.

F. Crop Retrieval

We show also asset retrievals based on cropped images. However, while the baselines might be adapted as layout estimators, it is still necessary to disentangle occlusions within and transform out environmental conditions, learning an invariance to lighting and occlusion (Fig. S.1).

G. Scene Retrieval

We show train-set images retrieved by CLIP similarity to test images in Fig. S.2. Identifying matches by image-level cosine similarity, we compute an LPIPS of 0.65, a CLIP similarity of 0.94, a DINOv2 similarity of 0.86, and a BioCLIP similarity of 0.58.

H. Asset Orientability

Preliminary to our investigation, we measure the ability of the visual encoders to orient the assets. Within each asset, we measure the pairwise cosine distance between each asset and each of its 72 rotations per five degrees, and take the mean across assets. Visualizations of this distance can be seen in Fig. S.6. The plots illustrate that only birds, carnivores, and herbivores are orientable, and that BioCLIP best distinguishes between orientations of the same asset. This matches our intuition that BioCLIP may be best-fitted as the target to our retrieval task since it is finetuned on taxonomic data to distinguish between species, but is not generally aligned with quantitative reconstruction results.

I. Additional Dataset Samples

We provide an additional 100 random dataset samples in Fig. S.7.

References

[1] Edward J. Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 1

[2] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1

[3] Peter Kulits, Haiwen Feng, Weiyang Liu, Victoria Fernandez Abrevaya, and Michael J. Black. Re-thinking inverse graphics with large language models. *TMLR*, 2024. 1

[4] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. YOLO-6D-Pose: Enhancing YOLO for single-stage monocular multi-object 6D pose estimation. In *3DV*, pages 1616–1625, Los Alamitos, CA, USA, 2024. IEEE, IEEE Computer Society. 1

[5] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE Press, 2020. 1

[6] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 1

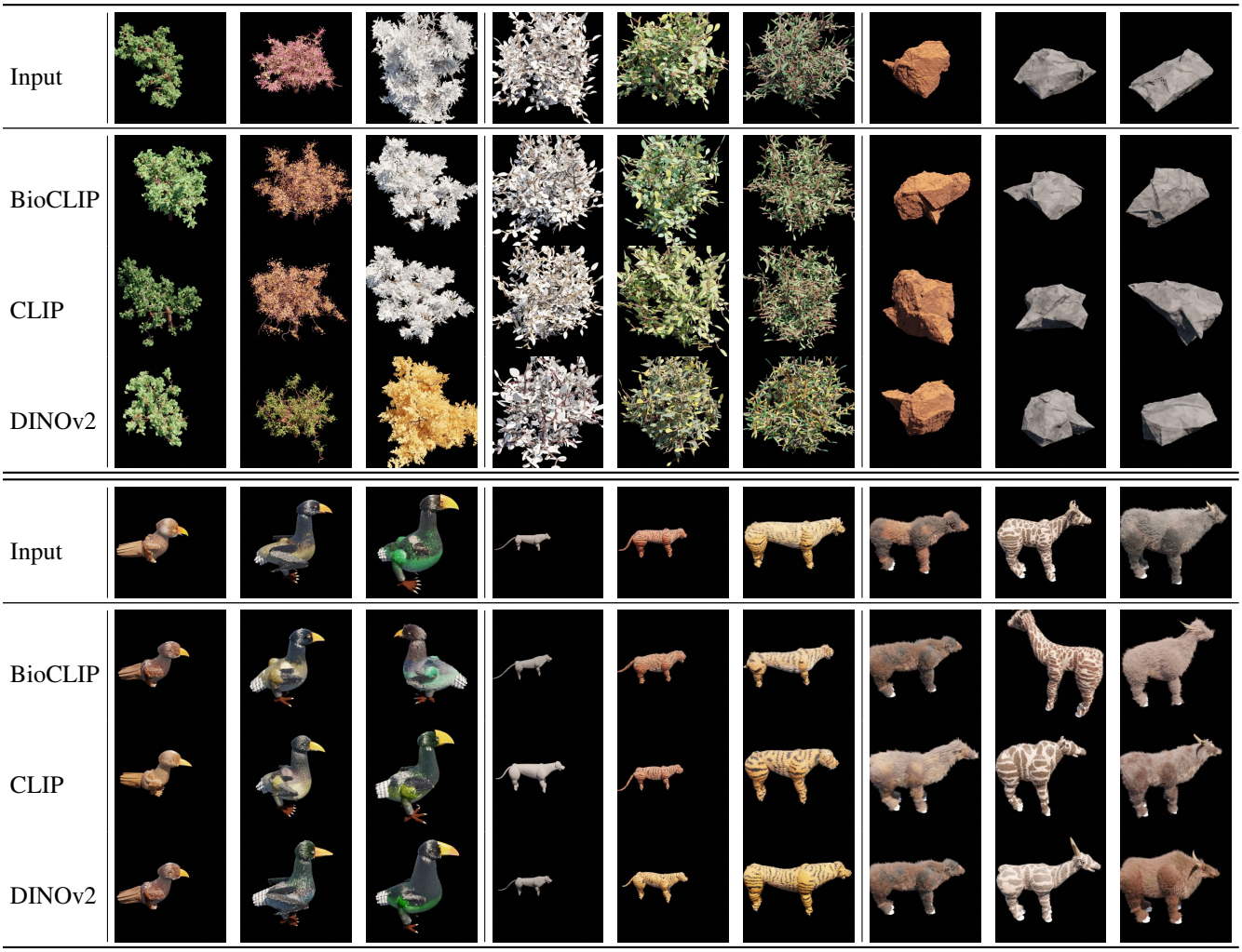

Figure S.4. **Retrieval Samples.** Database asset-retrieval samples from a held-out set of queries. Retrievals are from ground-truth embeddings and represent an upper bound on applied retrieval ability for each particular embedding type, as the embeddings are used as the target during training. Asset types from left to right: Top: tree, bush, boulder. Bottom: bird, carnivore, herbivore.

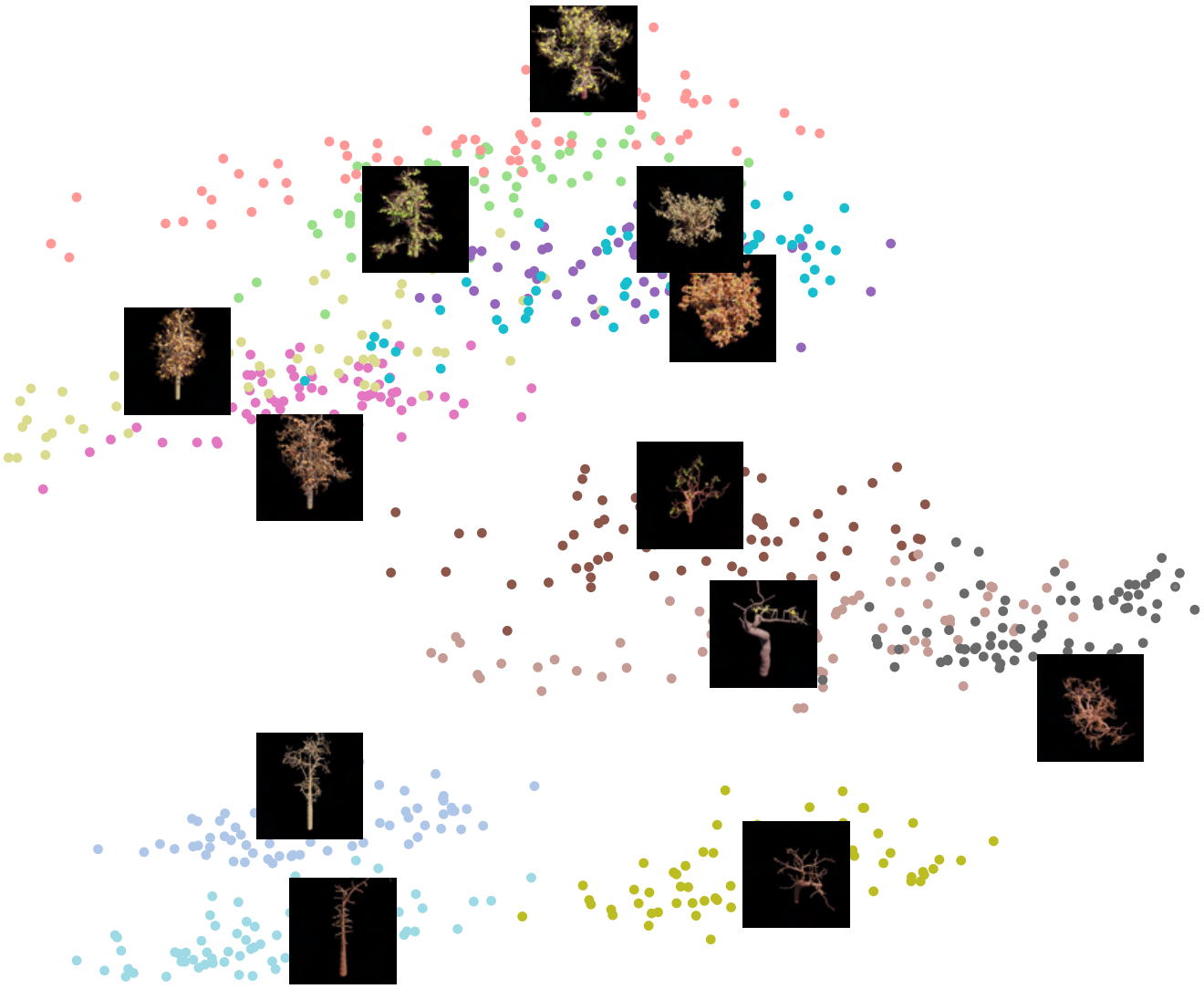

Figure S.5. **Asset-Space Visualization.** Visualization of the first two PCA components of the CLIP embeddings for a set of tree assets. Scatter dots with the same color represent five-degree rotations of a particular asset.

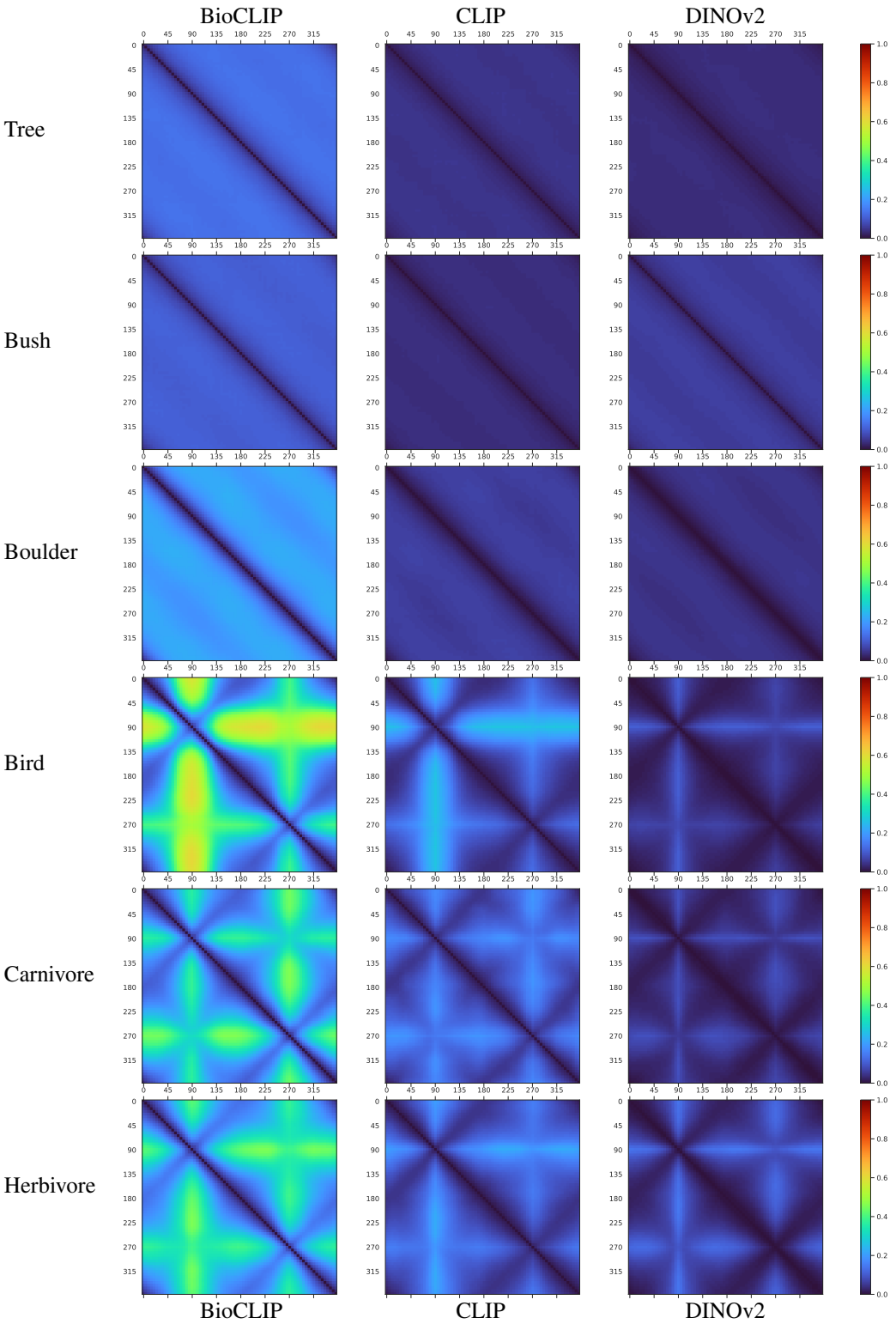

Figure S.6. **Asset Orientability.** Mean cosine distance between each pair of orientations, across assets. Greater values indicate that the visual encoder produces embeddings that better distinguish between assets at those pairs of orientations. We observe that only birds, carnivores, and herbivores are well oriented by the embeddings, that BioCLIP distinguishes most between orientations of the same asset, and that the differences appear small between all immediate pairs of rotations of each asset.

Figure S.7. **Additional Dataset Samples.**