Appendix

In this appendix, we further showcase the interpretability of ASIF models when used for classification in Figure 7. Then we provide additional details for the *scaling laws* and *EuroSAT* experiments presented in the main paper, and report additional results about the impact of the size of the encoders (Table 2), and of the image training dataset. Additionally, we briefly report an application of ASIF to a new modality (audio) in follow-up work by others. We also report further evidence that the ASIF construction is not overly sensitive to its hyperparameters. Lastly, we discuss more in detail the idea that captions of similar images are alike in Figure 10.

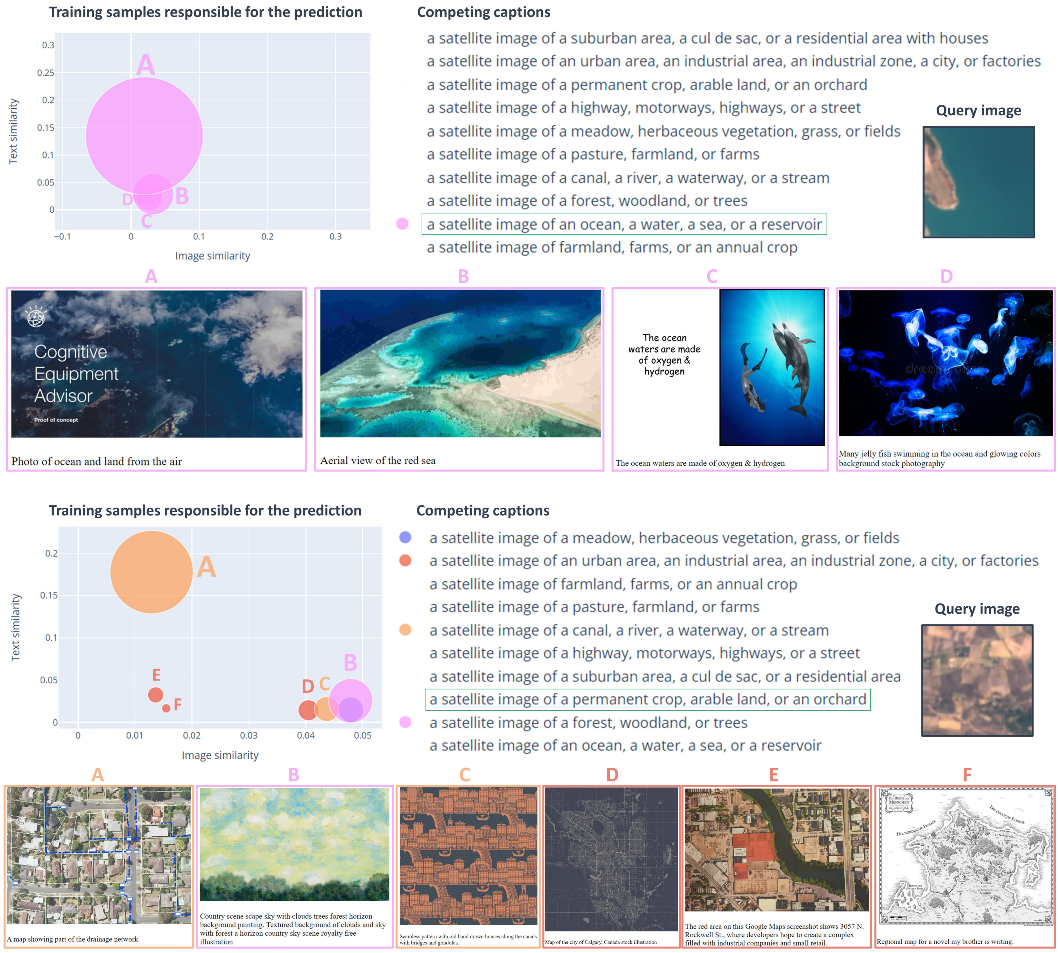

Figure 7: **Interpretability of EuroSAT classifications through ASIF.** Analysis of the classification outcome of two EuroSAT query images using ASIF. The scatter plot shows the samples in the training set closer to the query image and the candidate caption of the corresponding color. Image and text similarity are computed through cosine similarity in the visual space of DINO and the text space of SentenceT. The size of the marks is proportional to the product of the image and text similarity. The class chosen is the one with the largest total area. Below are shown the corresponding pairs from the training dataset CC12M. We can notice the distance between the EuroSAT dataset and the 1.6M samples of CC12M we used, many of the closest images are not from satellite and even then may have misleading descriptions, as image *A* in the second example. An interactive version of this plot for any ASIF classification can be obtained using our code demo attached in the supplementary material.

A Additional details on the scaling laws experiment

Models used in the scaling laws experiments. As discussed in the main paper, we tested ASIF with smaller image and text encoders to provide early evidence about ASIF scaling laws. We used three different instances of DEIT [43] vision transformers, the tiny (5.6M parameters, 192-dimensional embeddings), small (22M, 384), and base (87M, 768), and the original VITb16 vision transformer [55] (86M, 768). The DEIT models were pre-trained on a smaller dataset, the standard Imagenet1k training set [45], while VITb16 was pretrained on Imagenet21k [46]. As text encoders, we used smaller versions of SentenceT [47], with 23M and 33M parameters (both 384-dimensional embeddings), in contrast to the 110M parameters of the main model (768).

Figure 8 shows that, with smaller encoders producing smaller embedddings, we do not observe a performance saturation within 1.6M image-text couples. Further experiments with larger datasets are left for future work.

Impact of image pre-training data. In Table 2 we report the complete results of ASIF models using DEIT encoders [43]. We observe the expected positive correlation between the size of the encoders and the classification accuracy. Interestingly, ASIF with the largest instance of DEIT beats the one based on the standard VIT pretrained on Imagenet21k on three out of four of test datasets, while losing more than 10 points on CIFAR. These results may be interpreted in light of the similarity of the datasets we are using, with features useful to classify CIFAR images less overlapping with Imagenet1k features with respect to the other datasets.

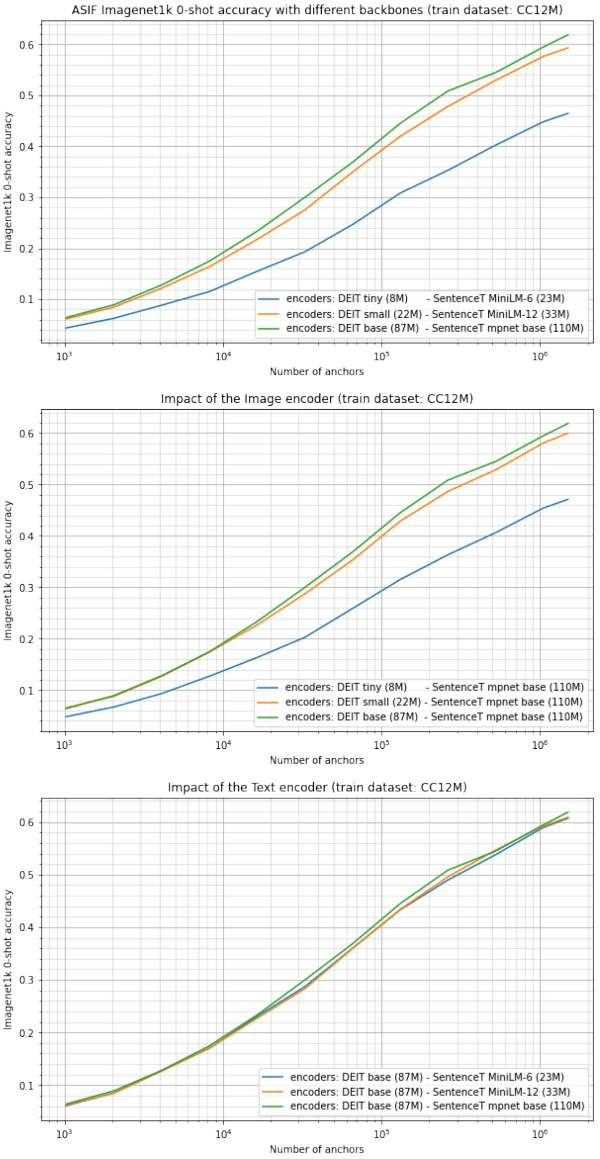

Figure 8: **ASIF performance does not saturate earlier with smaller encoders.** Classification accuracy keeps growing without saturating but is lower for smaller models. Furthermore, we observe that the quality of the vision encoder is more relevant than the quality of the text encoder with respect to zero-shot Imagenet classification.

B Additional details on the EuroSAT experiment.

EuroSAT, a renowned benchmark for satellite image classification, serves as a testing ground for out-of-distribution generalization in zero-shot and few-shot scenarios [52]. The dataset contains 27,000 images labeled under ten categories. Our ASIF model with a DINO visual backbone (denoted as 'ASIF unsup' in table 1) achieved a zero-shot classification score of 29.4%. While significantly better than random chance, this modest performance is not surprising considering the scarce presence of satellite images in the CC12M dataset.

As a further experiment, we randomly selected 100 images from the EuroSAT dataset and incorporated them into our ASIF training set, raising the total to 1,500,100 image-text pairs and leaving 26,900 images for testing. We created captions for the EuroSAT images using the template "*a satellite image of* [CLASS NAME]". This way the ASIF model improves dramatically, reaching a classification accuracy of $82.5 \pm 2.8\%$ on EuroSAT (average $\pm$ standard deviation of 5 trials).

ASIF backbones (Params, pre-training data)	**ImNet**	**CIFAR**	**PETS**	**ImNet-v2**
DEITtiny (5.6M, Im1k) - STminiL6 (23M, see Sec. 3)	46.5	37.3	75.6	38.3
DEITsmall (22M, Im1k) - STminiL12 (33M, see Sec. 3)	59.3	46.0	80.4	50.3
DEITbase (87M, Im1k) - STbase (110M, see Sec. 3)	60.9	50.2	81.5	52.2
VITb16 (86M, Im21k) - STbase (110M, see Sec. 3)	55.4	63.3	71.5	45.6

Table 2: **Zero shot classification accuracy of ASIF models with different backbones**. We observe that the ASIF procedure remains effective even with smaller encoders pre-trained on reduced visual datasets such as Imagenet1k.

Contrarily, CLIP [1], while demonstrating better zero-shot accuracy at 54.1%, is trained on a private dataset comprising 400 million images. This dataset may contain a larger number of satellite images than our 1.6 million subset of CC12M. Given the substantial improvement observed when we added just 100 EuroSAT images, it's reasonable to speculate that CLIP's enhanced performance might stem from its larger database of satellite images. However, confirming this theory is impossible due to the private nature of CLIP's training set.

We can, nevertheless, examine the presence of satellite images in the CC12M dataset. Using ASIF models' unique interpretability property, we can trace the training samples behind each classification. Figure 7 displays two EuroSAT samples, one classified correctly and the other not, along with the corresponding CC12M pairs responsible for the classifications. We note that our subset of CC12M is lacking in satellite images, and the few available often have misleading captions, such as a map of a drainage network tagged as "a satellite image of a canal, a river, a waterway, or a stream" instead of an urban area.

The images shown are an adaptation of the interactive plot to analyze any ASIF image classification we provided in the code demo attached in the supplementary material.

C ASIF used for audio in follow-up work.

Building on the work of ASIF, subsequent studies by other teams have not only adapted but also expanded its applications to encompass novel modalities, such as audio [*CITATION OMITTED for anonymity reasons, we report just their results in the inset, for the camera ready, we will replace this with the appropriate citation*].

The application of ASIF to audio has been primarily driven by its unique approach to retrieval through parallel anchors. In the context of speech-text representations, for example, ASIF's anchored retrieval allows probing the effectiveness of unimodal or non-unified spaces using paired multi-modal data, without further training. This quality becomes particularly noteworthy when direct cosine retrieval–a more traditional measure of similarity–is degraded.

Table 5: Text and speech encoder retrieval probe accuracy (%)

Method	LibriSpeech		AMI		CV	SWBD	TED
	test-clean	test-other	ihm	sdm1	test	test	test
LS Maestro (Direct)	20.5	19.3	7.65	6.16	7.43	13.88	11.89
LS Maestro (ASIF)	45.7	31.2	7.47	5.61	10.2	10.76	16.64
AMI Maestro (Direct)	67.2	48.9	45.2	32.6	19.0	44.7	43.9
AMI Maestro (ASIF)	33.6	17.7	14.9	10.5	7.88	16.7	21.5
CV Maestro (Direct)	76.3	61.7	19.1	10.0	40.0	29.3	44.8
CV Maestro (ASIF)	50.4	34.8	14.3	7.61	20.1	19.4	28.9
SWBD Maestro (Direct)	20.3	14.1	15.9	8.61	10.3	25.3	13.8
SWBD Maestro (ASIF)	49.0	23.0	13.8	7.32	8.94	19.7	29.0
TED Maestro (Direct)	80.6	64.3	24.5	13.3	29.0	40.6	77.9
TED Maestro (ASIF)	43.6	26.9	13.3	7.96	11.8	16.8	25.8
LS+C4 mSLAM (Direct)	1.96	2.0	1.54	1.10	1.5	1.63	1.52
LS+C4 mSLAM (ASIF)	8.63	10.5	3.99	3.06	1.79	6.03	5.78

The experiments on speech-text representations in this work have demonstrated that ASIF retrieval indeed exhibits improved performance over direct cosine retrieval in non-unified spaces (for example the ones produced by LS Maestro and SWBD Maestro as seen in the table in the inset). This observation validates the theoretical underpinnings of ASIF and its generalizability across varied modalities.

This acceptance and integration of ASIF into subsequent work highlights its value as a baseline for foundational multimodal models and underscores the significant role of retrieval methods in machine learning.

D ASIF sensibility to its hyperparameters

Finally, we present evidence about the sensitivity of the ASIF model to the hyperparameters p and k. Specifically, we show the hyperparameter search for PETS and CIFAR100 in Figure 9. Table D with results on the parameters fine-tuned on the two datasets reveals marginal improvements over the standard choice of k=800 and p=8. This suggests that the ASIF model is relatively insensitive to the choice of these hyperparameters.

Tuned on	**Parameters** p,k	**CIFAR**	**PETS**
PETS	(200,8)	60.9	72.3
CIFAR	(1600,6)	64.9	63
ImageNet1K	(800,8)	63.3	71.5

Table 3: Hyperparams search: tuning on each dataset per row.

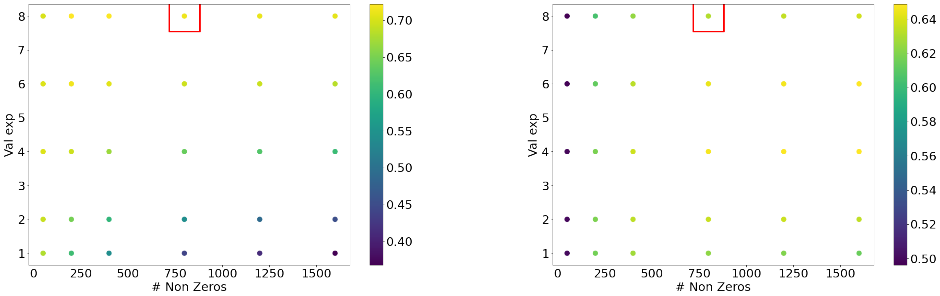

Figure 9: **Hyperparameters search** over Left Pets, Right CIFAR100. Highlighted in the red square the performance achieved tuning on Imagenet1K.

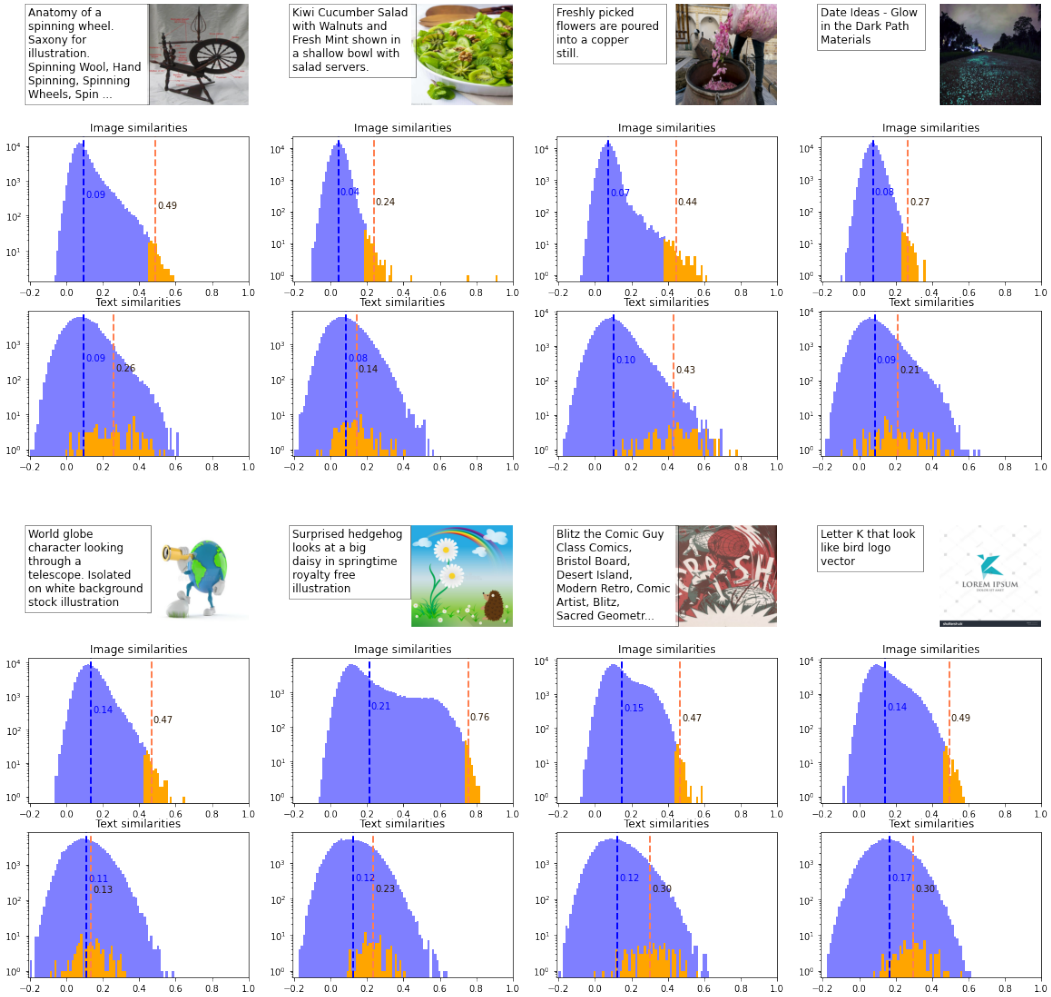

Figure 10: **Caption of similar images are themselves similar.** For 8 image-text pairs, we show in the first row the distribution of the image similarities against $100k$ images in the train set in blue (CC12M), and highlight the 1000 most similar in orange. The dashed lines indicate the mean of the two distributions. In the second row, we show the text similarities against the captions of the same $100k$ (blue) and 1000 (orange) images. If captions of similar images are themselves similar, we expect the dashed orange line in the second row to be at the right of the blue dashed line, as we observe. The average gap between the orange and blue lines in the second row over 10,000 image-text couples from CC12M is 0.098 ± 0.070.