

Efficient Fine-grained Image Retrieval with Vision Foundation Models for Industrial Objects

Supplementary Material

Table 4. Best binary image identification accuracy on 20k spare-part image pairs. A/WP denotes average/weighted-average pooling. All similarity scores are linearly normalized by logistic regression.

Train	Embeddings	Accuracy (%)	
Pre-trained	DINOv2[CLS]	91.33	
	DINOv3[CLS]	94.67	
	SigLIP2	94.26	
	DINOv2[P] + AP	72.77	
	DINOv3[P] + AP	62.33	
	DINOv2[CLS + P] + AP	89.36	
	DINOv3[CLS + P] + AP	93.63	
Linear Adaptation and Pooling	Global Features	SigLIP2	98.03
		DINOv2[CLS]	97.54
		DINOv3[CLS]	97.91
		DINOv2[CLS] + SigLIP2	97.79
		DINOv3[CLS] + SigLIP2	98.27
	+Local Features	DINOv2[P] + WP	97.46
		DINOv3[P] + WP	97.78
		DINOv2[P] + SigLIP2 + WP	97.78
		DINOv3[P] + SigLIP2 + WP	98.07
		DINOv2[CLS + P] + WP	97.00
		DINOv3[CLS + P] + WP	97.68
		DINOv2[CLS + P] + SigLIP2 + WP	97.84
		DINOv3[CLS + P] + SigLIP2 + WP	98.26

7. Best Performance of Image Identification

We additionally report the best performance across eight training seeds for each configuration to reflect the maximum discriminative potential that can be extracted from each embedding for this task. The results are consistent with the observations made in Section 5.1.

8. Additional Performance Comparisons of Object Retrieval

Here we present the performance difference between using CLS tokens and using patch tokens. The difference is defined by the retrieval score, i.e., the rank of gallery image in the retrieved images. Examples with smaller ranks are more interesting, as they correspond to more accurate retrievals, which are most relevant for real applications. Therefore, we define the difference between two retrieved rank as

$$d(r_1, r_2) = \frac{1}{\log(r_1 + 1)} - \frac{1}{\log(r_2 + 1)} \quad (8)$$

From top to bottom, the examples transit from cases where global representations (CLS tokens) perform better to cases where patch tokens perform better.

8.1. DINOv3[CLS] and DINOv3[P]

In Fig. 7, we show 24 samples ordered by the performance difference between DINOv3[CLS] and DINOv3[P] with weighted-average pooling. Beyond the observations in Sec. ??, we find that objects rarely seen during training (e.g., second-to-last and fourth-to-last rows) are poorly represented by the CLS token, and retrieved images are less relevant. This highlights a limitation in the generalization of DINOv3[CLS], as its global representation is less robust for underrepresented categories.

8.2. SigLIP2 + DINOv3[CLS] and SigLIP2 + DINOv3[CLS + P]

Image identification and retrieval results show that SigLIP2 CLS embeddings provide rich semantic representations and strong generalization. Combining SigLIP2 with DINOv3 reduces the performance gap among top-ranked retrievals, indicating that SigLIP2 stabilizes semantic features and mitigates discrepancies between CLS and patch representations.

In Fig. 8, 24 samples are ordered by the performance difference between SigLIP2 + DINOv3[CLS] and SigLIP2 + DINOv3[CLS + P]. In the last 12 examples, patch tokens improve performance by capturing 3D object structure, even in cases challenging for humans (e.g., third-to-last row). However, some inferred structures are incorrect, degrading retrieval. We attribute this to the lightweight linear pooling layer, which may be insufficient to fully exploit structural information in patch tokens—consistent with large transformer-based multi-view reconstruction models like VGGT [37] that directly use DINO patch tokens for geometry extraction.

9. Mini-Data

We provide a mini dataset of five objects, each with a query image, a random gallery, a challenge gallery, and a machine-generated image.



Figure 7. Retrieved results when using DINOv3[CLS] and DINOv3[P] + WP.

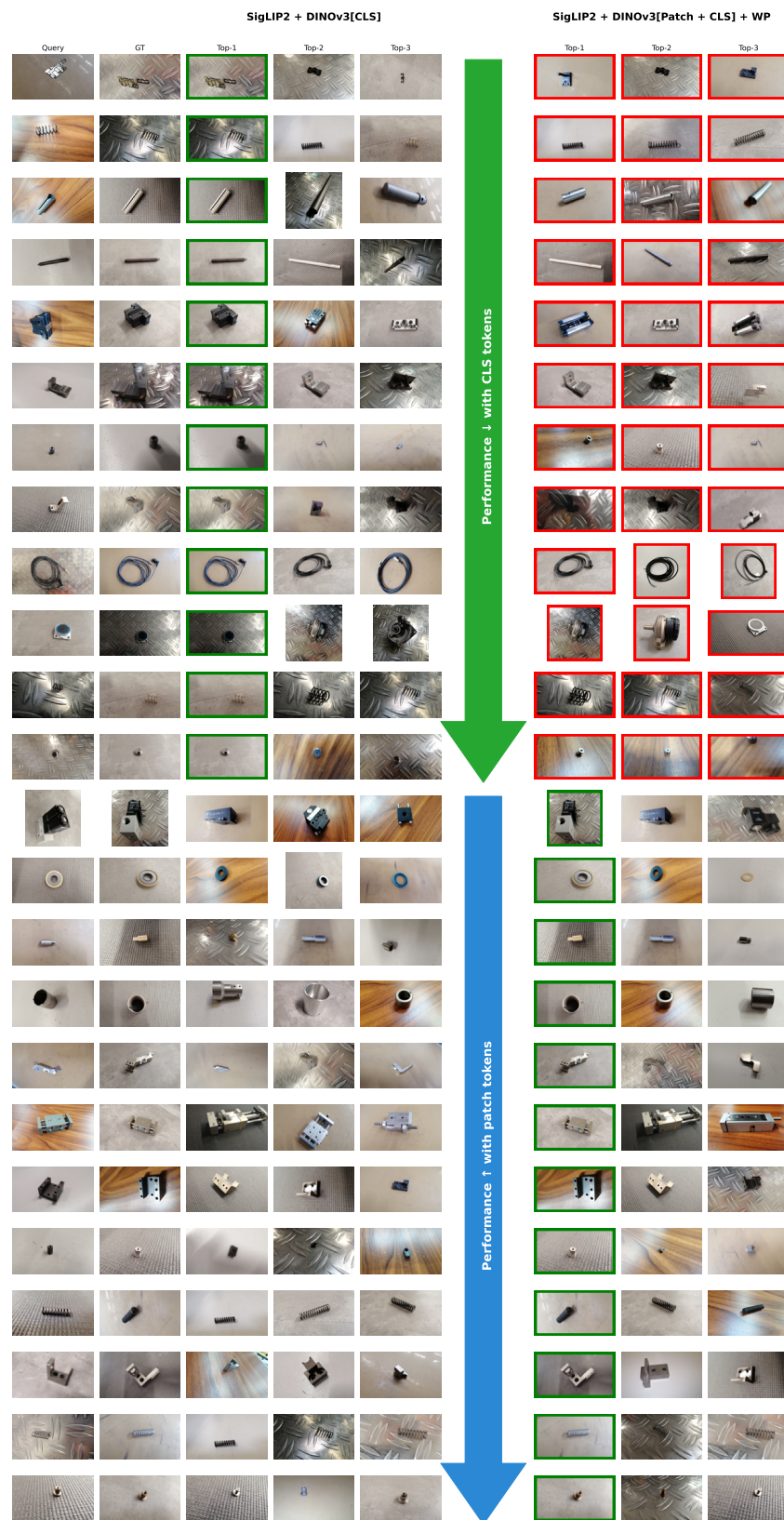


Figure 8. Retrieved results when using SigLIP2 + DINOv3[CLS] and SigLIP2 + DINOv3[CLS + P] + WP.