SearchAD: Large-Scale Rare Image Retrieval Dataset for Autonomous Driving

Supplementary Material

A. SearchAD Dataset

In this section, we provide more details on the SearchAD dataset (cf. Section 3) and its labeling project. In Figure 9, we showcase a cropped instance for each of the 90 semantic classes of Table 3, with all objects and scenes originating from the training split. Similar crops are also used for the vision support set of the search queries.

In Table 13, we illustrate the labeling guidelines for each of the nine categories using one representative class as an example. The complete labeling guidelines will be provided on the project page to maintain the scope of this supplementary material. Each entry provides a detailed label description, along with positive and negative examples. To clearly delineate relevant from irrelevant instances, the label description provides specific examples. A ball on a poster, for instance, is considered a false positive for the *Ball* class. Likewise, for *Fog*, strong rainy or snowy scenes are included as false positives, as their poor visibility results from conditions other than fog.

Furthermore, we maintained quality through weekly checks and by addressing label supplier questions, ensuring continuous adherence to our requirements. For challenging samples, we provided support to the label supplier upon request to determine whether the object or scene should be labeled. Moreover, this feedback also served as training material for the human experts.

Beyond the 90 official classes, SearchAD also provides catch-all classes for several categories, namely: *Animal-Real-Other*, *Animal-Statue-Other*, *Human-Duty-Other*, *Object-Movable-Other* and *Vehicle-Construction-Other*.

B. SearchAD Image Retrieval Benchmark

This section complements Section 4 by providing further details and more examples of the search queries, alongside a comparison of the test and validation splits.

Five more search queries, including their vision and language support sets, are illustrated in Table 10. This structure is standard for all remaining search queries within the SearchAD image retrieval benchmark, forming its default vision and language support sets. Specifically, language support sets contain up to three keywords and a total of four descriptions. The vision support set invariably consists of five images. Although some models might handle certain formulations better, we opted not to optimize the language support sets based on a specific subset of models. Instead, we defined them as precisely as possible, in alignment with the labeling guidelines. All 90 search queries, including their language and vision support sets are part of the benchmark and will be provided via the SearchAD devkit.

	Model	Test Split		Val Split	
		MAP [%]	MRP [%]	MAP [%]	MRP [%]
Text-to-Image	GDINO [29]	**5.25**	**6.49**	5.14	6.28
	OpenCLIP [24]	7.45	**10.17**	**7.90**	9.82
	SigLIP2 [50]	**8.57**	**11.24**	8.22	10.03
	BLIP2 [25]	9.14	**11.90**	**10.14**	11.77
	MetaCLIP2 [8]	**9.41**	**12.66**	8.74	11.23
	RADIO [20]	9.49	11.86	**10.44**	**12.18**
	NACLIP [19]	**9.59**	11.92	9.42	**12.28**
	NARADIO [1, 19]	**14.27**	**17.91**	13.22	15.60
Image-to-Image	OpenCLIP [24]	3.98	5.55	**4.16**	**5.75**
	RADIO [20]	4.34	6.00	**4.64**	**6.09**
	MetaCLIP2 [8]	**5.10**	**6.64**	4.49	5.64
	SigLIP2 [50]	**6.04**	**7.97**	5.42	7.39
	NACLIP [19]	6.56	9.30	**8.07**	**10.19**
	GDINO [29]	7.62	**10.45**	**8.41**	10.30
	BLIP2 [25]	7.95	10.82	**9.20**	**11.19**
	NARADIO [1, 19]	8.31	10.61	**10.04**	**12.10**

Table 7. Evaluation of retrieval methods on the SearchAD test and validation split. The table presents Mean R-Precision (MRP) and Mean Average Precision (MAP) for both splits, and for both text-to-image and image-to-image retrieval methods. Bold values highlight the top performance for each metric (MAP and MRP) within each row (model) across both splits. The results show that the test and validation splits are relatively well balanced.

In Table 7, we compare the results for all baseline methods on the validation split with those on the test split. Overall, the results are reasonably balanced between the two: 7 out of 16 methods show slightly better MAP performance on the test data, and 9 out of 16 methods exhibit slightly better MRP performance on this split. While text-based methods show a slight advantage on the test data, image-based methods exhibit a minor bias towards better performance on the validation split.

Text-to-Image Retrieval. Since the MAP for RADIO [20], BLIP2 [25], MetaCLIP2 [8] and NACLIP [19] all span from 9.14% up to 9.59%, their ranking varies in the validation split. While RADIO, MetaCLIP2, and BLIP2 successfully retrieved the image, including a group of firefighters, at the very top, achieving an AP of 100% for *Firefighter*, other models encountered difficulties with this challenging object. Notably, NACLIP [19] only achieved an AP of 3.33% for *Firefighter*, which allowed BLIP2 to even surpass NACLIP on the validation set. To better validate the somewhat underrepresented classes, the training split can be leveraged for zero-shot methods.

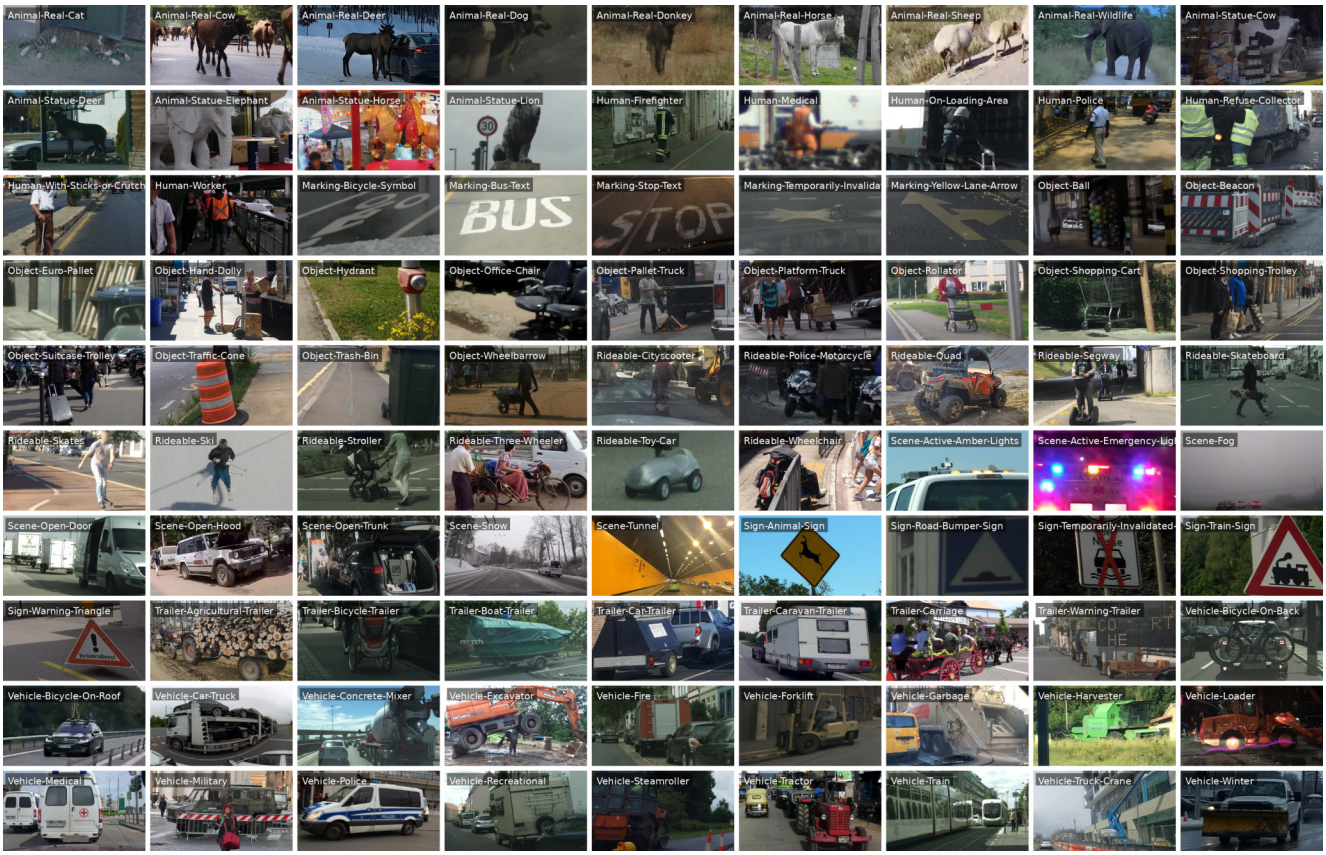


Figure 9. Overview of the 90 classes of SearchAD. Cropped instances of objects and scenes are provided for each class, extracted from the SearchAD training split. References for all underlying image sources are provided in Table 2.

Image-to-Image Retrieval. For the image-based methods, the ranking on the validation split is almost the same as the ranking on the test split. Nevertheless, MetaCLIP2 shows reduced performance on the validation set, primarily driven by a significant gap in AP for the two rideable vehicles *Police Motorcycle* and *Toy Car* between the validation and test set.

C. Detailed Baseline Results

The following section presents detailed class-wise baseline results, qualitative examples, and additional insights into the challenges GroundingDINO [29] encounters in text-to-image retrieval.

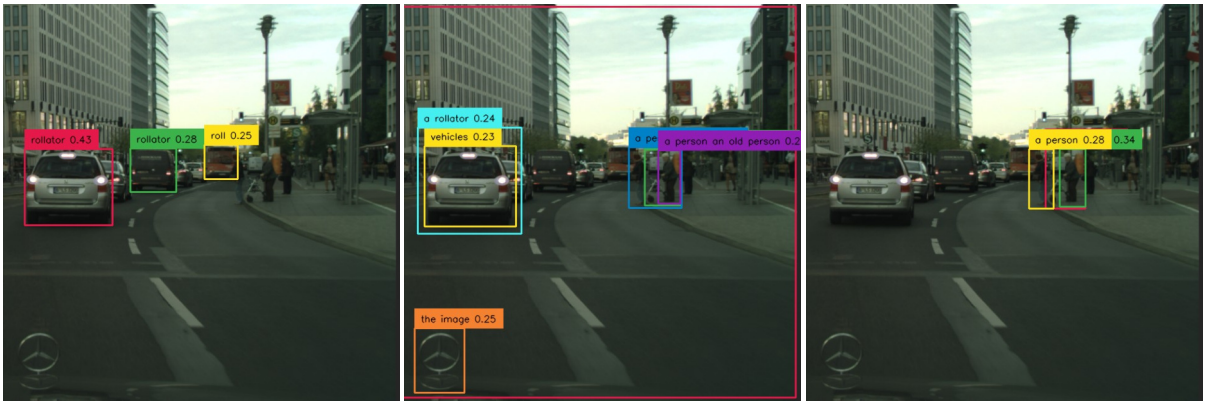
C.1. Class-wise Image Retrieval Results

Text-to-Image Retrieval. Table 11 shows the class-wise Average Precisions (APs) for text-based image retrieval across all baseline methods. The category-wise Mean Average Precisions (MAPs) can be derived by taking the mean of the respective class-wise AP scores, *e.g.*, for NARADIO [1, 19]:

$$\begin{aligned} \text{MAP}_{\textit{Marking}} &= \frac{31.45+16.12+9.72+0.99+7.88}{5} \\ &= 13.23\% \end{aligned}$$

The overall MAP is shown in the last row of the table and is calculated by directly averaging the AP scores across all classes. Table 11 demonstrates that NARADIO [1, 19] outperforms the other methods for 49 out of the 90 classes. However, the results also demonstrate that each model achieves the best performance for at least one class, underscoring the unique strengths and weaknesses of individual models and their pre-training. This highlights the importance of employing a diverse set of semantic classes for a comprehensive evaluation of zero-shot image retrieval capabilities.

Furthermore, the table reveals that more frequent classes, such as *Traffic Cone*, *Hydrant* or *Tunnel*, are generally easier to retrieve than rare classes, highlighting the key challenge of our proposed benchmark. However, not only rare, but also certain specific search queries, such as the scene classes *Open Door* and *Open Trunk*, also prove challenging. Even with well-described language support (*e.g.*, *Scene-*



(a) First Keyword: "Rollator" (b) Full Descriptions, cf. Table 10 (c) "A person walking with their rollator."

Figure 10. Visualization of the conflicting choice of a suitable text prompt for GroundingDINO [29]. The bounding box predictions for the class *Object-Rollator* are shown for different text inputs. Both box threshold and text threshold are set to 0.23 for better visualization. The search query of *Object-Rollator* along with its language support set can be found in Table 10.

Open-Trunk in Table 10), models struggle to accurately interpret and recognize their precise semantic meaning. Finally, models lacking text-aligned spatial features face significant difficulties with small objects. This is evident in the APs for the *Animal*, *Marking* and *Object* categories, where spatial and high-resolution (cf. Table 4) methods NARADIO and GroundingDINO [29] consistently outperform other methods.

Image-to-Image Retrieval. In Table 12, the class-wise results for the image-based methods are shown. In general, text-to-image retrieval methods demonstrate superior performance compared to image-to-image retrieval. Image-to-image retrieval outperforms or matches the best text-to-image method for only 29 out of 90 classes. This superior performance does not appear to follow a clear categorical pattern, instead occurring somewhat randomly across various classes. Notably, for the two classes *Train* and *Police Motorcycle*, the best performing image-to-image method achieved an AP more than 10% higher than the best performing text-to-image retrieval method.

For image-to-image retrieval, NARADIO again emerges as the overall best performing model and leads in many individual classes. However, unlike in text-to-image retrieval, other methods such as BLIP2 [25] and GroundingDINO show comparable strong performance. For example, GroundingDINO achieves the best results for 8 out of the 15 *Object* classes. Conversely, the image-based OpenCLIP [24] method do not achieve the best performance for any of the 90 classes.

C.2. Qualitative Image Retrieval Results

Based on the retrieval results on the validation split of SearchAD, we showcase the top 5 ranked images of all eight text-to-image retrieval methods. The corresponding search queries, along with their language support sets, can be found in Table 10. Notably, some retrieved images might appear quite similar, as certain datasets include temporally correlated recordings (e.g., video frames rather than individual keyframes).

Figure 12 illustrates the top retrieval results when searching for *Tractors*. The qualitative results confirm the MAP scores shown in Table 11, demonstrating that models with text-aligned spatial features generally achieve superior results. Notably, GroundingDINO [29] struggles to retrieve the correct vehicles in this context. OpenCLIP [24] incorrectly ranks a semantically similar *harvester* as second, an image that should ideally be retrieved when searching for the very rare *Vehicle-Harvester* class. While all models exhibit at least one false positive image within their top 5 results, NARADIO [1, 19] is the only method to achieve a perfect precision of 100% within its top 5 results (Precision@5 [34]).

However, Figure 13 and Figure 14 highlight a persistent challenge: all models struggle to retrieve precisely described semantic classes. The results highlight a common failure mode where models correctly identify the semantic class but also retrieve images from incorrect categories (cf. Figure 8). Specifically, all eight methods in Figure 13 incorrectly retrieve at least one image featuring a traffic sign symbolizing a cow. For *Marking-Bus-Text* in Figure 14, all models except for NARDIO and NACLIP [19] tend to retrieve generic road markings right next to the ego vehicle instead of the specific 'BUS' text. Addressing this issue by better conditioning specific categories remains an area for future research.

The results in Figure 15 highlight another limitation: Models struggle to completely comprehend the semantic descriptions for specific search queries. For instance, when searching for scenes like *Open Trunk*, the presence of a large trunk seems to be prioritized over the critical detail of

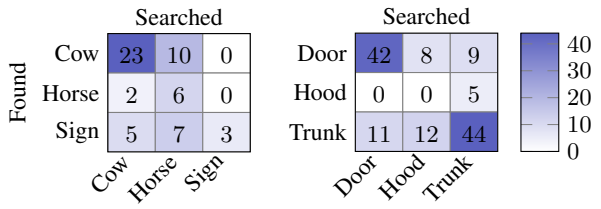


Figure 11. Retrieval confusion matrices based on the Top-50 results of NARADIO. Queries: (i) Animal-Real-Cow, Animal-Real-Horse, Sign-Animal-Sign and (ii) Scene-Open-Door, Scene-Open-Hood, Scene-Open-Trunk.

it being open. While most models struggle to retrieve any correct images within their top 5 results, NARADIO once again achieves a Precision@5 of 100%.

C.3. Text Prompt Sensitivity in GroundingDINO

We chose GroundingDINO [29] as one of our baseline methods due to its excellent ability to recognize open-world objects, which represent some of the most challenging retrieval instances in our benchmark. As detailed in Table 11, GroundingDINO also demonstrates the best performance across all models mainly for object and animal classes. GroundingDINO excels particularly with classic and simple object and animal classes, such as *Dog*, *Horse*, *Hydrant*, or *Stroller*.

However, its performance significantly degrades for other classes. This decline is primarily attributed to its sensitivity to text prompt choices, as exemplified by the *Object-Rollator* class in Figure 10. While GroundingDINO generally achieves its best results with precise and short descriptions, issues arise with less optimal prompts. For instance, when using only the first keyword, as shown in Figure 10a, "Rollator" is often unknown to the model, resulting in a similarity score below the threshold of 0.23 for the object-of-interest. Meanwhile, the car in front is detected with a high similarity score of 0.43, leading to numerous false positives in the top retrieval results.

Conversely, employing a comprehensive description can also introduce false positives. For example, the Mercedes-Benz star might be erroneously detected with the label "the image" (similarity score of 0.25), and the car in front of the ego vehicle is again misidentified as a rollator. For the *Rollator* class, the most effective prompt was found to be "A person walking with their rollator." which directs the model's attention more appropriately towards pedestrians.

In contrast, many other VLM-based text-to-image retrieval methods show significantly greater robustness to prompt engineering as they achieve a similar overall MAP regardless of whether using only keywords or both keywords and descriptions. This robustness, combined with the ability to pre-calculate and store image features in optimized search indices like FAISS [12], makes VLM-based methods considerably more suitable for real-time search and practical applications demanding rapid results.

	Model - BLIP2 [25]	MAP[%]	MRP[%]
Language	BLIP2 - Baseline	9.14	11.90
	BLIP2 - Max. Query	6.20	8.06
	BLIP2 - Eval Small	1.45	1.62
	BLIP2 - Eval Large	13.51	15.75
Vision	BLIP2 - Baseline	7.95	10.82
	BLIP2 - Random Sets	5.36 ± 0.23	7.20 ± 0.22
	BLIP2 - Max. Query	5.73	7.88
	BLIP2 - RPN	12.46	16.15

Table 8. Ablation studies based on SearchAD test set.

C.4. Quantitative Failure Mode Analysis

We extend our NARADIO [1, 19] results and quantify failure modes via Top-50 confusion matrices (Figure 11). The analysis reveals semantic ambiguity (e.g., *Cow* queries retrieve *Animal Signs* more frequently than explicit sign queries do) and visual similarity (e.g., *Open Trunk* retrieves *Open Hoods*). Derived from the confusion matrices, the highest Precision@50 achieved is 88% for *Open Trunk*.

D. Ablation Studies based on BLIP2

D.1. Ablations of Vision and Language Support Sets

Queries were optimized for superior validation accuracy across all baselines, without model-specific support set optimization. Leveraging insights from labeling QCs, vision support images were manually selected for high variance, representativeness, and low occlusion. We validated our curation against 20 random selections from the training set (cf. Table 8 - Baseline vs. Random Sets). Mean queries outperform maximal similarity across single queries (cf. Table 8 - Baseline vs. Max. Query), highlighting their robustness against confusing correlations.

D.2. Class-Agnostic Region Proposal Preprocessing

While our main paper intentionally established plain baselines, we validated incorporating Faster R-CNN's RPN [44] as a pre-selection stage for plausible region proposals. As a second step, each region will be processed separately by BLIP2 [25]. Unlike refined detectors, which suppress unknown classes (e.g., road markings) as background, we opt for raw RPN proposals to maintain high recall for rare objects. Augmenting these proposals with the entire image (crucial for scenes like *Fog*) yields a substantial MAP boost from 7.95% to 12.46% (cf. Table 8). This confirms that region-based processing is effective, though more an engineering solution.

D.3. Size-Dependent Evaluation

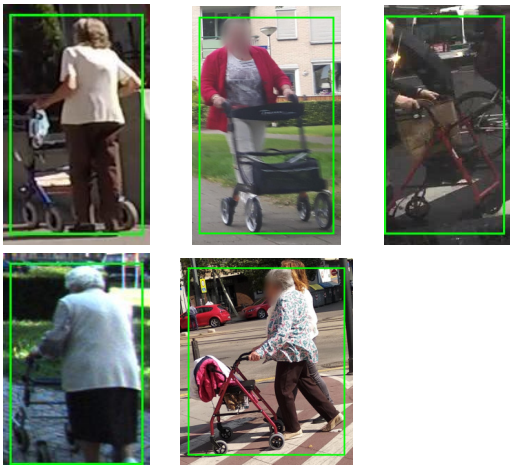
We isolated object size as a critical performance bottleneck: As show in Table 8 MAP collapses from 13.51% (largest 33% of objects) to 1.45% (smallest 33%), quantitatively confirming the challenge of small object retrieval.

Search Time Optimization	t_{Search}	t_{Setup}	MAP [%]
BLIP2 - No Search Index [25]	5.55s	179.29s	**9.14**
BLIP2 - Faiss Index [12]	**0.28s**	**28.23s**	8.74
GroundingDINO [29]	21 - 23h	-	5.25

Table 9. MAP and search time for VLM-based methods (with and without index) and grounded object detectors on SearchAD Test.

D.4. Optimized Search Index

Table 9 compares both retrieval accuracy and runtime for VLM-based methods based on BLIP2 [25] with and without index and grounded object detectors. The results are based on the entire SearchAD test set. While the MAP only reduces by 0.4%, both search time and setup time are significantly reduced by the index. When scaling the dataset to millions of images, the index will be inevitable to enable AD developers fast search across AD databases. Moreover, the search time of more than 20 hours for GroundingDINO [29] demonstrates that grounded object detectors are generally not suitable for real-world applications.

Class	Language Support Set	Vision Support Set
Animal-Real-Cow	Keywords: "Cow", "Cattle", "Bovine", Descriptions: "A cow next to the street or standing on the street.", "The cow is standing in front of the vehicle or below the vehicle.", "There is a cow in the driving scene on the road or on the sidewalk.", "There are cows next to the road on a pasture."	
Marking-Bus-Text	Keywords: "Bus Road Marking", "Bus Text Pattern", Descriptions: "The word bus is painted on the street in white color.", "The word bus is painted on the street in yellow color.", "The bus text pattern on the road indicates the bus lane.", "The image contains a bus road marking that may indicate a bus stop."	
Object-Rollator	Keywords: "Rollator", "Walking Aid", "Mobility Walker" Descriptions: "A person walking with their rollator.", "A rollator is visible in the image, requiring vehicles to be aware of potential pedestrians with mobility issues.", "The image shows a walking aid, indicating a need for increased awareness of vulnerable road users.", "There is an old person using a rollator for mobility assistance."	

Continued on next page

Table 10 – Continued from previous page

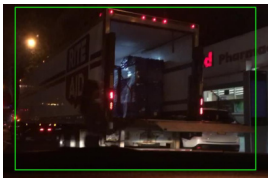
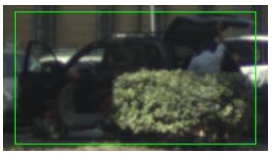
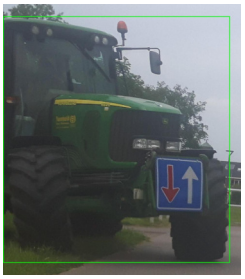
Class	Language Support Set	Vision Support Set
Scene-Open-Trunk	Keywords: "Open Car Trunk", "Vehicle with open trunk.", "Van or truck with open trunk doors.", Descriptions: "The image shows a passenger car, van or truck with an open trunk.", "The van has its trunk doors to the back of the vehicle open.", "The image contains a vehicle with an open trunk.", "The open trunk of the vehicle allows to look into the trunk or the loading area of the truck."	    
Vehicle-Tractor	Keywords: "Tractor", "Farm Tractor", "Agricultural Tractor", Descriptions: "A tractor is a versatile vehicle designed to deliver a high tractive effort at slow speeds, used for plowing, planting, and other farm tasks.", "A tractor is visible in the image, requiring vehicles to be aware of its presence on rural roads and fields.", "The image contains a tractor on the country road or on the field in the background next to the road.", "The image shows a large green or red agricultural tractor with a front shovel."	    

Table 10. Illustration of five search queries from different categories including their vision and language support sets.

Class	Class-wise Average Precision (AP) [%]							
	GDINO [29]	OpenCLIP [24]	SigLIP2 [50]	BLIP2 [25]	MetaCLIP2 [8]	RADIO [20]	NACLIP [19]	NARADIO [1, 19]
Animal-Real-Cat	0.11	0.18	0.09	0.03	0.03	0.07	0.11	**0.14**
Animal-Real-Cow	16.86	7.97	16.55	9.22	14.21	13.40	27.96	**28.25**
Animal-Real-Deer	**48.04**	6.67	6.69	39.40	33.53	38.49	10.43	40.90
Animal-Real-Dog	**29.74**	8.95	13.71	12.78	16.06	10.67	20.39	28.23
Animal-Real-Donkey	0.00	0.01	0.07	0.02	0.03	0.15	**0.22**	0.10
Animal-Real-Horse	**14.92**	1.31	1.36	2.40	0.97	3.76	3.44	4.83
Animal-Real-Sheep	10.71	3.41	17.66	10.42	10.47	12.82	11.87	**28.97**
Animal-Real-Wildlife	0.02	0.01	0.19	0.01	0.08	1.57	**16.70**	4.22
Animal-Statue-Cow	0.00	0.00	0.00	**0.02**	0.00	0.00	0.00	0.01
Animal-Statue-Deer	33.34	33.34	6.69	33.34	33.34	**33.36**	33.34	33.34
Animal-Statue-Elephant	0.01	0.02	0.06	0.07	0.01	0.03	0.08	**0.27**
Animal-Statue-Horse	0.17	0.33	1.41	1.19	0.49	1.04	0.87	**6.59**
Animal-Statue-Lion	0.36	2.30	1.83	4.29	4.13	1.99	5.90	**13.93**
Human-Construction-Worker	1.18	12.91	16.27	15.64	16.57	14.45	13.14	**18.12**
Human-Firefighter	0.03	26.59	7.66	27.39	6.94	**41.53**	4.98	7.34
Human-Medical	0.01	0.14	0.04	0.00	0.04	0.00	3.85	**7.15**
Human-On-Loading-Area	0.41	1.69	2.10	1.19	1.47	1.67	0.73	**2.51**
Human-Police	1.39	1.90	8.35	6.07	**18.68**	12.52	7.56	15.56
Human-Refuse-Collector	0.04	7.63	10.18	11.04	**18.16**	11.44	3.92	0.95
Human-With-Sticks-or-Crutches	1.07	0.82	5.56	1.64	0.84	0.80	2.33	**16.56**
Marking-Bicycle-Symbol	4.29	12.06	10.21	16.91	10.92	7.78	19.14	**31.45**
Marking-Bus-Text	1.67	3.31	4.74	6.28	5.76	4.39	11.52	**16.12**
Marking-Stop-Text	0.53	2.36	2.50	5.33	2.05	2.87	3.10	**9.72**
Marking-Temporarily-Invalidated	0.03	0.06	0.42	0.10	0.05	0.14	0.07	**0.99**
Marking-Yellow-Lane-Arrow	0.74	0.72	0.82	1.94	0.78	1.06	1.27	**7.88**
Object-Ball	14.69	3.09	11.41	3.98	5.36	6.89	3.98	**21.69**
Object-Beacon	5.21	5.76	8.04	7.89	**12.00**	9.09	6.28	8.38
Object-Euro-Pallet	0.60	4.07	12.66	3.95	19.10	13.72	9.29	**24.98**
Object-Hand-Dolly	0.19	2.49	5.10	1.15	**10.58**	9.62	1.53	5.87
Object-Hydrant	**43.25**	14.82	10.08	13.15	14.73	16.23	13.21	20.97
Object-Office-Chair	0.21	0.02	0.01	0.01	0.01	0.03	0.01	**0.04**
Object-Pallet-Truck	0.19	0.40	7.17	0.46	2.31	**7.19**	0.34	6.75
Object-Platform-Truck	0.33	0.46	0.55	0.48	0.94	**1.69**	0.33	0.51
Object-Rollator	0.10	0.26	0.52	0.40	2.68	0.29	0.89	**10.85**
Object-Shopping-Cart	0.55	6.42	1.21	3.69	2.21	3.02	6.31	**6.65**
Object-Shopping-Trolley	0.57	0.53	0.60	0.91	0.54	0.44	1.16	**2.29**
Object-Suitcase-Trolley	3.90	1.30	1.72	2.26	2.11	2.16	2.28	**6.79**
Object-Traffic-Cone	**60.69**	22.34	22.42	28.62	26.94	28.26	45.75	59.11
Object-Trash-Bin	10.77	10.47	16.00	8.32	12.27	8.44	12.18	**29.48**
Object-Wheelbarrow	1.10	4.43	0.75	**10.33**	0.87	0.66	6.61	7.31
Rideable-Cityscooter	0.29	0.72	1.06	0.48	2.49	1.00	0.94	**4.44**
Rideable-Police-Motorcycle	0.58	**44.78**	7.87	34.78	10.69	10.85	8.41	5.86
Rideable-Quad	0.20	10.89	11.05	11.05	10.99	11.34	10.88	**13.00**
Rideable-Segway	10.47	0.61	0.03	0.03	13.92	0.05	0.38	**14.95**
Rideable-Skateboard	2.28	2.45	3.92	2.38	3.07	2.27	**4.90**	4.05
Rideable-Skates	0.34	2.01	1.26	2.25	4.01	20.02	1.44	**33.41**
Rideable-Ski	0.31	1.21	8.34	0.22	0.51	17.69	**22.50**	5.50
Rideable-Stroller	**30.63**	5.37	5.78	6.18	4.16	5.65	8.48	18.99
Rideable-Three-Wheeler	17.80	9.99	14.55	24.85	**36.59**	26.89	12.07	27.04
Rideable-Toy-Car	0.65	13.63	**41.64**	6.98	34.53	26.77	9.06	33.61
Rideable-Wheelchair	2.31	1.28	2.96	1.40	1.69	3.28	1.48	**5.65**
Scene-Active-Amber-Lights	0.12	0.51	0.38	0.81	0.83	**1.21**	0.43	0.80
Scene-Active-Emergency-Lights	0.10	0.23	0.49	**2.11**	0.22	1.35	0.77	1.93

Table 11 – Continued on next page

Table 11 – Continued from previous page

Class	Class-wise Average Precision (AP) [%]							
	GDINO [29]	OpenCLIP [24]	SigLIP2 [50]	BLIP2 [25]	MetaCLIP2 [8]	RADIO [20]	NACLIP [19]	NARADIO [1, 19]
Scene-Fog	0.87	86.36	71.19	**87.33**	68.92	61.45	86.86	1.58
Scene-Open-Door	1.71	2.26	3.22	2.47	2.70	2.62	3.00	**7.58**
Scene-Open-Hood	0.07	0.09	**6.86**	3.00	0.59	0.35	0.06	0.25
Scene-Open-Trunk	1.85	2.93	4.29	3.20	3.57	3.30	2.70	**7.58**
Scene-Snow	6.63	60.34	54.60	66.47	56.87	70.17	**74.79**	61.49
Scene-Tunnel	2.74	33.25	24.11	32.46	20.13	38.29	39.01	**41.71**
Sign-Animal-Sign	1.41	2.35	4.72	3.37	**6.15**	3.91	2.96	3.66
Sign-Road-Bumper-Sign	1.10	1.94	2.44	1.51	2.04	1.96	1.35	**2.55**
Sign-Temporarily-Invalidated-Sign	0.05	0.10	0.12	**0.15**	0.09	0.10	0.11	0.13
Sign-Train-Sign	0.62	1.87	**3.82**	2.11	1.57	1.87	1.51	1.34
Sign-Warning-Triangle	0.17	0.14	6.40	0.31	3.70	0.21	1.02	**12.93**
Trailer-Agricultural-Trailer	0.17	0.33	0.50	0.71	**1.40**	1.34	0.56	**1.40**
Trailer-Bicycle-Trailer	0.11	0.15	0.88	1.14	0.83	0.41	0.50	**4.20**
Trailer-Boat-Trailer	0.10	12.41	7.86	0.53	3.25	0.47	6.77	**28.22**
Trailer-Car-Trailer	4.59	2.44	4.45	5.25	4.87	5.15	6.46	**18.21**
Trailer-Caravan-Trailer	0.32	2.57	1.59	2.11	2.88	1.57	4.42	**11.52**
Trailer-Carriage	16.19	8.09	20.22	18.01	15.16	12.52	9.92	**46.44**
Trailer-Warning-Trailer	0.47	0.92	1.78	1.13	**3.01**	1.85	1.00	2.11
Vehicle-Concrete-Mixer	3.23	18.80	26.23	20.50	25.29	18.49	24.30	**32.84**
Vehicle-Excavator	7.05	7.65	14.91	12.84	16.97	14.20	23.02	**31.37**
Vehicle-Forklift	3.02	6.11	8.86	4.52	2.34	5.55	**15.15**	15.03
Vehicle-Harvester	0.01	0.91	0.04	0.12	0.01	0.04	**0.75**	0.34
Vehicle-Loader	0.38	3.38	3.08	5.75	4.90	5.33	9.36	**19.26**
Vehicle-Steamroller	0.18	1.44	4.18	2.26	5.31	6.28	4.38	**15.85**
Vehicle-Tractor	0.59	6.97	8.61	5.68	8.34	5.54	11.16	**25.01**
Vehicle-Truck-Crane	3.65	9.62	11.89	10.21	11.65	11.05	15.45	**19.14**
Vehicle-Fire	4.97	14.81	12.77	**16.94**	12.75	12.32	6.00	10.98
Vehicle-Garbage	1.59	16.79	17.84	13.04	**21.21**	9.17	19.62	5.38
Vehicle-Medical	6.58	14.25	17.46	16.28	15.84	8.14	14.48	**18.45**
Vehicle-Military	0.05	4.25	27.98	24.90	**29.89**	30.12	2.70	16.83
Vehicle-Police	0.57	2.67	4.70	6.89	4.98	6.70	4.76	**27.49**
Vehicle-Winter	0.01	6.78	0.37	1.10	0.60	1.46	**13.10**	0.26
Vehicle-Bicycle-On-Back	0.16	1.26	7.43	2.18	2.40	3.08	0.33	**11.41**
Vehicle-Bicycle-On-Roof	0.02	0.01	**0.19**	0.02	0.01	0.02	0.01	0.05
Vehicle-Car-Truck	0.07	0.82	6.70	11.56	**11.94**	9.60	0.91	10.57
Vehicle-Recreational	1.16	5.83	9.75	14.45	24.48	15.86	22.51	**28.12**
Vehicle-Train	25.79	24.03	36.72	36.48	18.20	31.35	36.90	**38.35**
Mean over all classes (MAP)	5.25	7.45	8.57	9.14	9.41	9.49	9.59	**14.27**

Table 11. Evaluation of text-to-image retrieval methods on the SearchAD test set: Class-wise Average Precision (AP), and the Mean Average Precision (MAP) over all 90 classes. Bold values indicate the highest scores across the models. The models (columns) are ordered ascending by their overall MAP.

Class	Class-wise Average Precision (AP) [%]							
	OpenCLIP [24]	RADIO [20]	MetaCLIP2 [8]	SigLIP2 [50]	NACLIP [19]	GDINO [29]	BLIP2 [25]	NARADIO [1, 19]
Animal-Real-Cat	0.02	0.02	0.01	0.02	0.03	**0.07**	0.06	0.02
Animal-Real-Cow	0.39	0.19	2.19	0.25	**17.63**	10.51	13.06	0.26
Animal-Real-Deer	33.39	33.48	33.62	33.34	**69.70**	21.11	33.57	33.35
Animal-Real-Dog	5.02	1.47	2.11	4.70	18.85	**20.73**	9.16	6.91
Animal-Real-Donkey	0.01	0.03	0.02	0.01	0.01	**0.05**	0.00	0.01
Animal-Real-Horse	0.03	0.03	0.02	0.02	**4.27**	2.28	0.57	0.10
Animal-Real-Sheep	0.04	0.13	0.03	0.05	**11.20**	0.96	4.47	3.80
Animal-Real-Wildlife	0.04	0.02	0.00	0.01	0.11	**5.57**	0.00	0.00
Animal-Statue-Cow	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Animal-Statue-Deer	33.34	33.35	33.34	33.34	**33.36**	33.34	33.35	33.34
Animal-Statue-Elephant	0.02	1.24	0.01	0.08	0.01	0.26	0.01	**0.08**
Animal-Statue-Horse	0.40	2.77	2.93	**3.00**	2.52	1.93	0.56	1.39
Animal-Statue-Lion	0.05	0.05	0.05	0.05	**8.39**	3.73	1.93	0.07
Human-Construction-Worker	1.38	2.79	1.92	2.38	**11.68**	9.77	8.94	9.70
Human-Firefighter	4.57	0.02	0.03	0.03	7.48	0.11	**27.14**	0.05
Human-Medical	0.01	0.00	0.01	0.00	**0.02**	0.01	0.00	0.00
Human-On-Loading-Area	0.62	0.45	0.55	0.50	**2.11**	0.95	1.61	0.46
Human-Police	0.78	2.44	7.57	0.84	**17.00**	6.76	14.77	15.55
Human-Refuse-Collector	0.73	0.54	0.58	**10.75**	2.59	1.19	9.03	0.27
Human-With-Sticks-or-Crutches	0.74	0.45	0.64	0.80	0.73	0.68	**0.96**	0.55
Marking-Bicycle-Symbol	5.98	4.24	5.07	6.12	8.16	**17.42**	14.63	5.57
Marking-Bus-Text	2.85	1.71	3.41	2.53	4.46	6.57	**7.94**	5.40
Marking-Stop-Text	1.15	0.88	0.56	0.70	0.44	0.92	**3.46**	1.44
Marking-Temporarily-Invalidated	0.10	0.04	0.11	0.06	0.18	**0.32**	0.16	0.06
Marking-Yellow-Lane-Arrow	0.52	0.56	0.47	0.47	4.45	**4.59**	1.80	0.92
Object-Ball	0.15	0.09	0.14	0.41	0.09	**22.52**	1.86	0.93
Object-Beacon	6.45	7.94	7.46	12.02	17.57	**25.95**	25.93	20.13
Object-Euro-Pallet	5.00	3.04	4.20	**19.90**	4.36	2.62	12.63	9.73
Object-Hand-Dolly	1.44	0.33	0.34	1.11	1.86	**2.09**	0.38	0.39
Object-Hydrant	7.19	10.54	8.67	10.54	8.10	**13.27**	7.87	8.10
Object-Office-Chair	0.01	0.04	0.01	0.01	0.00	**0.40**	0.01	0.01
Object-Pallet-Truck	0.36	0.52	0.39	**0.60**	0.20	0.08	0.54	0.36
Object-Platform-Truck	0.38	0.58	0.41	0.38	**1.21**	0.57	0.44	0.36
Object-Rollator	0.22	0.11	0.18	0.22	1.07	**1.24**	0.44	0.61
Object-Shopping-Cart	0.61	3.44	1.16	0.20	6.79	5.73	1.71	**9.02**
Object-Shopping-Trolley	0.60	**1.06**	0.41	0.77	0.83	0.62	0.53	0.93
Object-Suitcase-Trolley	2.12	2.13	2.60	2.82	2.12	2.75	2.90	**5.46**
Object-Traffic-Cone	10.68	13.77	10.89	13.28	36.88	**51.95**	27.12	41.18
Object-Trash-Bin	12.44	5.89	11.12	11.75	6.75	**19.78**	14.86	16.13
Object-Wheelbarrow	0.59	0.08	0.31	0.31	**13.39**	2.70	2.20	0.09
Rideable-Cityscooter	0.38	0.31	0.63	0.29	1.29	0.46	**0.78**	0.37
Rideable-Police-Motorcycle	2.73	0.07	**59.10**	0.64	13.11	4.98	34.95	0.33
Rideable-Quad	2.70	0.05	9.72	**11.02**	10.94	8.32	7.21	0.08
Rideable-Segway	0.58	0.01	0.48	**8.75**	0.02	3.32	0.04	0.05
Rideable-Skateboard	0.58	0.20	1.94	**2.24**	0.12	0.16	0.30	0.81
Rideable-Skates	0.03	0.92	0.01	0.43	0.03	0.51	**0.97**	0.06
Rideable-Ski	6.65	**25.10**	0.17	12.51	0.10	0.05	8.34	21.67
Rideable-Stroller	2.00	0.93	2.31	1.75	11.06	**12.09**	4.08	1.31
Rideable-Three-Wheeler	1.44	5.91	1.38	1.87	19.47	20.89	7.92	**21.33**
Rideable-Toy-Car	2.35	1.06	1.75	**24.20**	2.91	1.45	9.64	2.48
Rideable-Wheelchair	0.51	1.14	0.17	1.66	2.26	1.91	1.12	**8.67**
Scene-Active-Amber-Lights	0.24	1.13	0.37	0.60	0.78	0.60	1.14	**3.05**
Scene-Active-Emergency-Lights	0.11	0.13	0.17	0.13	0.11	0.16	**0.30**	0.15

Table 12 – Continued on next page

Table 12 – Continued from previous page

Class	Class-wise Average Precision (AP) [%]							
	OpenCLIP [24]	RADIO [20]	MetaCLIP2 [8]	SigLIP2 [50]	NACLIP [19]	GDINO [29]	BLIP2 [25]	NARADIO [1, 19]
Scene-Fog	71.57	23.94	46.04	53.21	2.79	**77.07**	70.22	58.32
Scene-Open-Door	2.50	2.04	2.12	2.73	1.83	**2.96**	2.47	2.53
Scene-Open-Hood	0.15	0.12	0.08	**0.20**	0.08	0.08	0.11	0.10
Scene-Open-Trunk	2.98	2.15	2.52	2.81	2.50	3.16	**3.20**	2.20
Scene-Snow	10.58	34.12	7.32	9.98	7.43	9.95	**41.99**	11.07
Scene-Tunnel	13.18	20.07	19.02	15.69	4.32	4.55	**26.92**	5.87
Sign-Animal-Sign	1.13	**3.48**	0.55	1.34	2.96	1.24	1.41	2.79
Sign-Road-Bumper-Sign	1.08	1.07	0.94	1.08	1.03	**5.66**	2.06	2.73
Sign-Temporarily-Invalidated-Sign	0.17	0.11	0.11	0.15	0.11	**0.84**	0.20	0.31
Sign-Train-Sign	1.01	0.59	0.90	1.20	1.57	**2.02**	1.60	0.64
Sign-Warning-Triangle	0.25	0.50	0.38	**5.87**	0.92	2.00	0.74	2.55
Trailer-Agricultural-Trailer	4.31	3.22	7.29	**9.39**	2.05	0.52	3.31	2.12
Trailer-Bicycle-Trailer	0.24	0.11	**2.86**	0.36	0.28	0.24	0.70	3.01
Trailer-Boat-Trailer	0.66	0.09	3.55	8.63	18.68	5.98	7.12	**31.67**
Trailer-Car-Trailer	3.89	2.42	**6.60**	4.06	1.88	3.96	4.33	6.53
Trailer-Caravan-Trailer	0.59	0.91	0.69	0.85	1.77	3.66	2.23	**6.90**
Trailer-Carriage	7.64	6.44	8.64	13.45	13.78	34.38	20.63	**42.04**
Trailer-Warning-Trailer	3.41	4.31	**4.85**	3.76	0.96	1.24	2.51	2.99
Vehicle-Concrete-Mixer	2.80	6.30	13.51	8.74	10.20	**24.30**	16.05	23.86
Vehicle-Excavator	2.09	6.84	2.40	8.27	9.32	16.14	10.79	**31.34**
Vehicle-Forklift	0.44	6.34	0.53	6.42	9.33	5.26	12.46	**13.66**
Vehicle-Harvester	0.03	0.03	0.02	0.03	0.27	**0.38**	0.13	0.29
Vehicle-Loader	2.27	4.20	1.07	2.74	12.15	**17.32**	5.16	17.54
Vehicle-Steamroller	0.29	3.61	0.25	0.75	5.33	**10.86**	3.49	15.66
Vehicle-Tractor	3.05	2.67	4.17	8.82	7.16	6.61	7.84	**25.15**
Vehicle-Truck-Crane	7.20	10.06	3.28	9.65	6.03	4.64	11.06	**14.83**
Vehicle-Fire	2.20	4.05	4.78	2.98	8.26	9.35	**9.67**	2.44
Vehicle-Garbage	7.31	2.95	**14.32**	14.04	2.48	3.49	6.09	8.83
Vehicle-Medical	5.26	5.35	5.14	7.47	3.08	3.99	5.24	**22.92**
Vehicle-Military	6.07	1.11	30.85	**35.16**	0.92	2.36	18.96	0.80
Vehicle-Police	4.40	3.74	3.96	6.14	1.76	9.03	6.84	**27.24**
Vehicle-Winter	0.41	0.18	0.11	**2.15**	0.20	0.15	0.48	0.42
Vehicle-Bicycle-On-Back	0.31	1.30	0.22	**2.77**	0.37	0.23	0.22	0.91
Vehicle-Bicycle-On-Roof	0.01	0.02	0.02	0.03	0.07	0.02	0.03	**0.08**
Vehicle-Car-Truck	0.91	8.06	9.98	**10.94**	2.17	5.62	6.65	1.74
Vehicle-Recreational	9.73	7.69	9.94	**11.07**	9.13	6.67	6.33	10.34
Vehicle-Train	17.34	37.19	18.34	29.91	48.81	42.90	32.84	**52.66**
Mean over all classes (MAP)	3.98	4.34	5.10	6.04	6.56	7.62	7.95	**8.31**

Table 12. Evaluation of image-to-image retrieval methods on the SearchAD test set: Class-wise Average Precision (AP), and the Mean Average Precision (MAP) over all 90 classes. Bold values indicate the highest scores across the models. The models (columns) are ordered ascending by their overall MAP.

Figure 12. Text-based retrieval results on the validation split for the *Vehicle-Tractor* class, showing top 5 ranked images for each model. Model references can be found in Table 11.

Figure 13. Text-based retrieval results on the validation split for the *Animal-Real-Cow* class, showing top 5 ranked images for each model. Model references can be found in Table 11.

Figure 14. Text-based retrieval results on the validation split for the *Marking-Bus-Text* class, showing top 5 ranked images for each model. Model references can be found in Table 11.

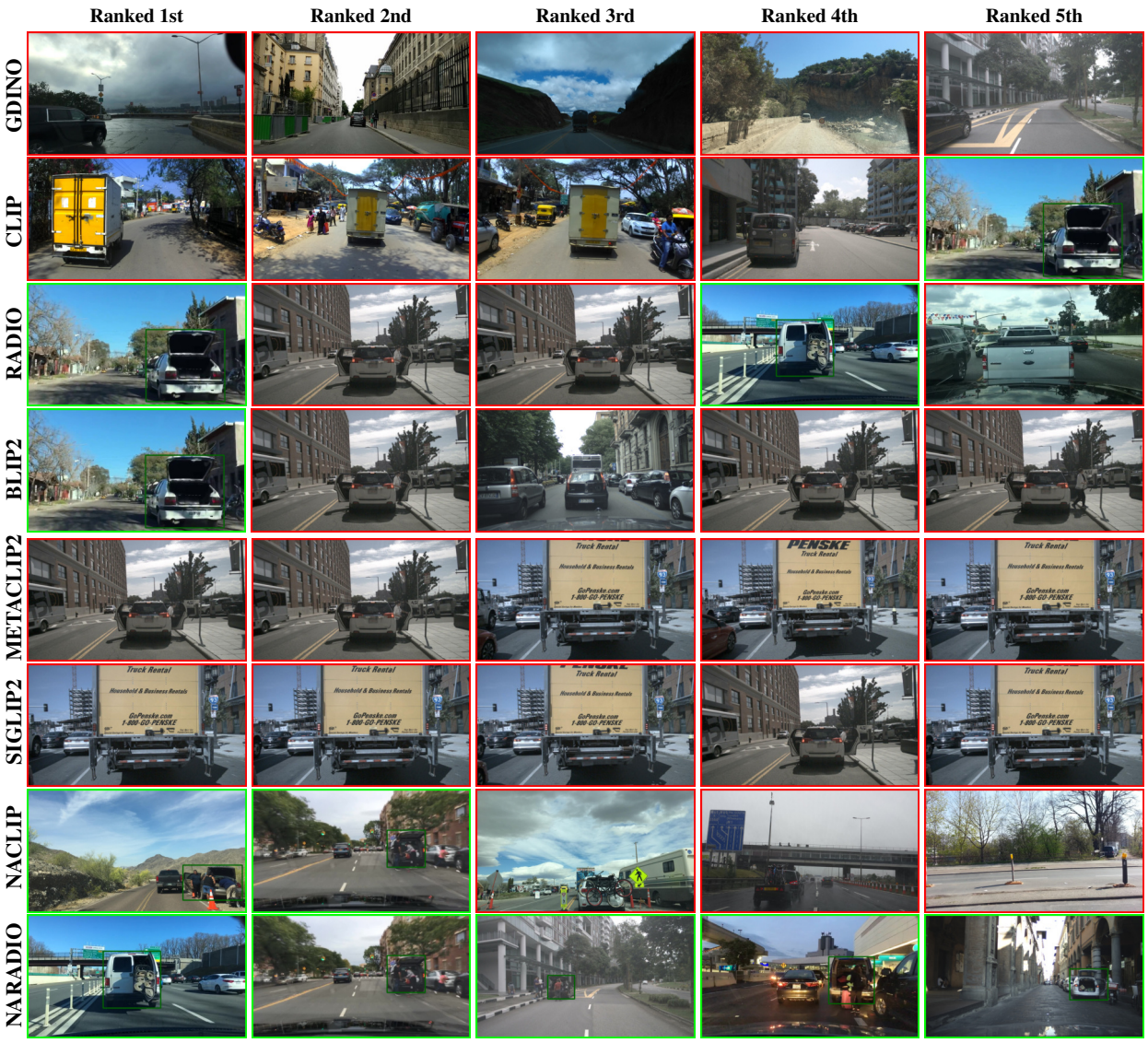


Figure 15. Text-based retrieval results on the validation split for the *Scene-Open-Trunk* class, showing top 5 ranked images for each model. Model references can be found in Table 11.

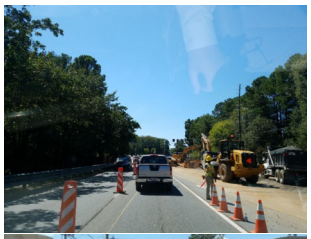
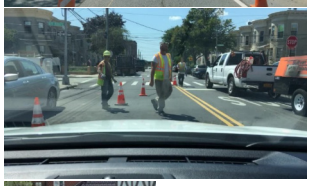
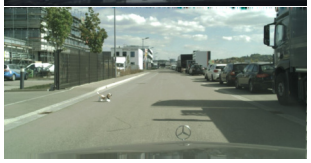
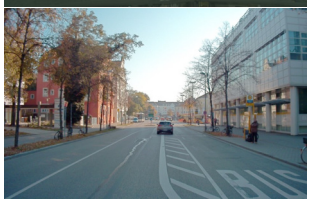
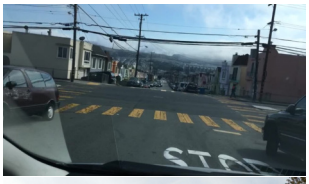
Class	Label Description	Positive Examples	Negative Examples
Human-Construction-Worker	• Ideal case: Person with yellow vest and helmet, standing on a construction site • Person with helmet • Person with warning / high-visibility vest (vest does not have to be yellow or orange) • To not label the person inside a (construction) vehicle. Only when it is a rideable vehicle or when the door is open and he/she is getting in or out • People standing within the construction site or moving materials should also be labeled as Pedestrian Duty Construction. • Hard negative example: Police officer or fire fighter with warning vest, other duty worker (*e.g.*, garbage collector)	  	
Animal-Real-Dog	• All kinds of dogs • Artificial dog *e.g.*, dog-dummy / puppet (*e.g.*, in Lost and Found [40] dataset) that is indistinguishable from a real dog should also be counted as a dog • Any 3D dog dummies of dogs should be also counted as dog dataset) • Hard negative example: a dog printed on a poster (2D dog)	 	
Marking-Bus-Text	• Find the text pattern “BUS” on the road • The text pattern must be on the road, not on a traffic sign or somewhere else • The text pattern “BUS” can be in any color • It can be occluded, but it has to be clear that it is the “BUS” text pattern, *e.g.*, if it is identifiable by the context • Label based on context — for example, if there’s a nearby bus station or a bus stop sign, it should be annotated. • Hard negative example: Any other text pattern, which looks similar to the “BUS” text pattern	 	 

Table 13 – *Continued on next page*

Table 13 – Continued from previous page

Class	Label Description	Positive Examples	Negative Examples
Object-Ball	• All kinds of balls (football ball, tennis ball, basketball ball, . . .) • The ball must be real (in 3D) and not shown on a poster • Bounding box should include the person carrying or playing the ball if there is one. Otherwise label only the ball itself • Hard negative examples: Ball printed on a poster, football field, where you cannot see a football, helmets, any other spherical objects, *e.g.*, round street lamps		
Rideable-Wheelchair	• Wheelchair usually has four wheels and is designed for people who cannot walk • Also if there is no person on the wheelchair • Usually the wheelchair is pushed by someone but some also have a motor • Bounding box should include the person in the wheelchair if there is one. Otherwise label only the wheelchair itself • Hard positive example: Wheelchair without any person on the sidewalk • Hard negative example: Rollator, Shopping trolley, Stroller, Shopping cart		
Scene-Fog	• Definition of fog: When visibility is less than one kilometer • Otherwise it is only haze • Hard negative example: Poor visibility due to heavy rain • Choose a large part of the image where you can also see the effect of the fog and use this as bounding box (see example)		

Table 13 – *Continued on next page*

Table 13 – Continued from previous page

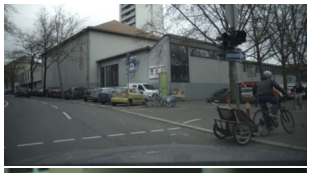
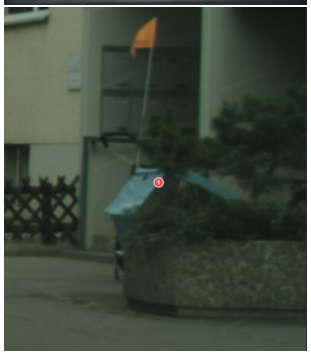
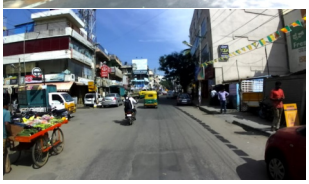
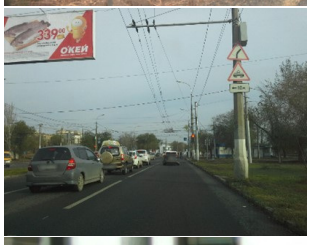
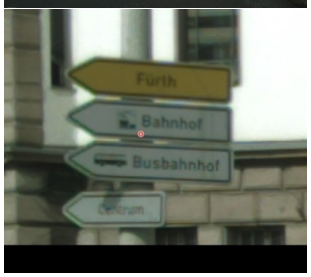
Class	Label Description	Positive Examples	Negative Examples
Trailer-Bicycle-Trailer	• Bicycle trailer for children • Bicycle trailer to transport things • Also if there is no bicycle in the front (*e.g.*, if the bicycle trailer stands on the sidewalk) • Trailer must be intended for bicycles/E-bikes and not for vehicles and motorcycles • Bicycle trailer has to be a trailer and no cargo bike (see negative example)	 	  
Sign-Train-Sign	• There has to be a train symbol on the traffic sign • Symbols could be a modern train, steam train or a tram • Only traffic signs and no real trains • Hard negative examples: Any other traffic signs that warn of a railway crossing, *e.g.*, Andrew cross	  	 

Table 13 – *Continued on next page*

Table 13 – Continued from previous page

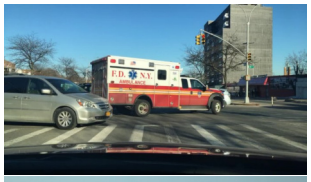
Class	Label Description	Positive Examples	Negative Examples
Vehicle-Medical	• All kinds of ambulance vehicles, usually red/orange and white colored • Ambulances can usually be identified by their text lettering (*e.g.*, Ambulance, Ambulancia, Krankenwagen, Emergency Medical Service, American Red Cross, Rettungsdienst, Rotes Kreuz) • Some vehicles are both police and ambulance vehicles • Some vehicles come from both the fire department and the ambulance • Hard negative example: Small fire truck	  	

Table 13. SearchAD labeling guidelines: For each category, we provide one example class to showcase the labeling guidelines. All positive and negative images are extracted from the eleven SearchAD datasets [4, 6, 9, 13, 15, 36, 40, 46, 51, 58, 59].