# SOCO: Benchmarking Semantic Object Correspondence in Vision Foundation Models

Olaf Dünkel[1⋆], Basavaraj Sunagad[2⋆], Haoran Wang[1], David T. Hoffmann[3], Christian Theobalt[1], and Adam Kortylewski[2]

[1] Max Planck Institute for Informatics, SIC
[2] CISPA Helmholtz Center for Information Security
[3] University of Freiburg

**Project Page:** https://genintel.github.io/SOCO/

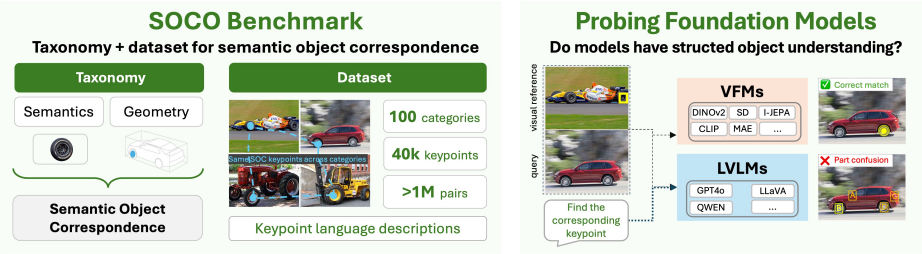



**Fig. 1:** SOCO provides the first taxonomy-driven, language-grounded formulation of **Semantic Object Correspondence** (SOC), enabling structured, semantically coherent, and cross-category part annotations across 100 diverse categories, which allows evaluating semantic and structured object understanding in vision foundation models (VFMs) and large vision language models (LVLMs).

**Abstract.** Measuring structured object understanding in vision foundation models remains challenging due to inconsistent evaluation protocols and limited part-level supervision. Semantic correspondence (SC) evaluates this capability by testing whether object parts can be matched across instances and categories under large variations in appearance, viewpoint, and geometry. To enable a systematic SC evaluation, we introduce SOCO, a new benchmark for Semantic Object Correspondence that introduces a taxonomy of correspondence types and provides consistent, functionally meaningful keypoint annotations across 100 categories and over 1M correspondence pairs. In addition, SOCO includes keypoint language descriptions, enabling the evaluation of large vision-language models (LVLMs) and their fine-grained part-level understanding. Comprehensive experiments reveal that (i) vision foundation backbones encode strong semantic structure but transfer correspondences poorly across related categories and only partially capture object-part position, (ii) LVLMs are stronger at text-prompted part localization than at visual-reference cross-image matching, exposing a gap between language-grounded localization and fine-grained visual correspondence, and (iii) correspondence performance predicts dense downstream tasks—segmentation, tracking, 3D pose estimation, and 3D detection—more

⋆ Equal contribution.

strongly than ImageNet classification. Together, these findings position SOCO as a benchmark for structured, part-level representation quality in vision and multimodal foundation models.

**Keywords:** Semantic Correspondence · Representation Learning · Benchmarking

## 1 Introduction

Visual representations form the foundation of visual intelligence. Evaluating their quality has long been central to progress in computer vision. Existing benchmarks probe distinct aspects of visual understanding, ranging from category-level recognition benchmarks such as ImageNet [16] to spatial localization tasks including detection, segmentation, and pose estimation [2,15,36]. However, they provide limited insight into whether a representation captures *structured object understanding*, i.e., the ability to relate semantically corresponding parts across different object instances and categories.

Recently, *semantic correspondence (SC)* has become increasingly important for evaluating self-supervised and foundation models [24,51,60,66], as it measures a model's ability to establish correspondences between object parts across different instances of a category—a capability that requires consistently capturing object structure under substantial variation in appearance, viewpoint, and geometry. The ability to establish such correspondences is crucial for transferring knowledge across related objects, for example when adapting affordances, recognition, pose estimation, or reconstruction to unseen categories, which is important for embodied and robotic systems [23,68].

However, despite this growing adoption, progress in SC research has been constrained by the lack of a clear task definition and by the limitations of existing datasets [45,63]. Current benchmarks conflate *two distinct abilities* in a single within-category score—recognizing the same local concept (e.g. a wheel center) and identifying its correct *repeated instance* within an object (front-left vs. rear-right wheel)—and do not evaluate transfer *across related categories* (a wheel center on a car, bus, or tractor) at all. This ambiguity limits current evaluations of modern foundation models.

We therefore propose **Semantic Object Correspondence (SOC)**, a taxonomy-driven formulation of semantic correspondence that disentangles these three abilities. SOC explicitly models the relationship between object part semantics and the overall object structure, providing a clearer separation between local concept recognition, object-relative identity, and cross-category transfer. Concretely, the taxonomy distinguishes *concept correspondence* (CC, matching the same local concept), *semantic object correspondence* (SOC, matching the same concept with the same object-relative identity), and *cross-category SOC* (matching object-relative keypoints across related categories through shared taxonomy concepts). This decomposition reduces annotation ambiguity, standardizes what constitutes a valid correspondence across object categories and viewpoints, and makes distinct model failure modes separately measurable.

Building on this definition, we present **SOCO**, a **S**emantic **O**bject **CO**rrespondence dataset that measures SOC with taxonomy-driven keypoint annotations across *100 object categories* organized into four super-classes. Unlike prior datasets, SOCO emphasizes semantic consistency and cross-category matching, enabling structured evaluation of correspondence across varying geometry and appearance for a diverse range of man-made object and animal categories. Across a broad family of vision foundation models—including self-supervised and vision–language models such as DINO [13,47,60], CLIP [48], Stable Diffusion [53], and I-JEPA [3]—the decomposed evaluation reveals distinct failure modes: strong VFMs recognize local concepts but exhibit large CC→SOC drops (repeated-part confusion) and further SOC→Cross-SOC drops (limited category-level abstraction). Moreover, SOC is a practical zero-shot diagnostic of representation quality: it correlates with dense downstream tasks—segmentation, tracking, 3D pose, and 3D detection—more strongly than ImageNet classification accuracy.

As connecting vision and language modalities becomes increasingly important, benchmarks for structured object understanding should not only evaluate visual representations but also *large vision–language models (LVLMs)*. To support this, we extend SOCO with *language descriptions of correspondence keypoints*, creating a comprehensive benchmark for studying the interplay between visual correspondences and natural language in multimodal foundation models. LVLM evaluations reveal a complementary failure mode: current LVLMs are substantially stronger at text-prompted part localization within a single image than at visual-reference correspondence across images, exposing a gap between language-grounded localization and visual matching. Together, the results position SOCO as a unified benchmark for analyzing fine-grained visual reasoning and multimodal representation quality in the era of large foundation models.

**In summary, our main contributions are:**

- **Task formulation.** We introduce **Semantic Object Correspondence (SOC)** as a taxonomy-driven decomposition of semantic correspondence into concept correspondence, structured object understanding, and cross-category transfer.
- **Dataset.** We present **SOCO**, a large-scale benchmark built on this taxonomy, featuring 100 diverse categories, semantically grounded keypoint annotations, and over 1M correspondence pairs, with provided **language descriptions** that enable joint study of visual correspondence and language understanding in multimodal models.
- **Vision-model analysis.** Across a broad family of vision foundation models, the SOC decomposition exposes repeated-part confusion and limited cross-category abstraction even in strong dense self-supervised backbones.
- **LVLM analysis.** On the same taxonomy, current LVLMs are stronger at text-prompted part localization than at visual-reference cross-image matching, revealing a gap between language-grounded localization and fine-grained visual correspondence.

**Table 1: Comparison of semantic correspondence benchmarks.** SOCO uniquely combines a hierarchical keypoint taxonomy, language descriptions, and cross-category correspondence pairs while covering a large and diverse set of categories, compared to other SC datasets that include man-made objects.

| Dataset | #Cats. | #Pairs | Keyp. | Taxonomy | Sep.geo. | Cross-cat. | Language |
|---|---|---|---|---|---|---|---|
| PF-WILLOW | 5 | 900 | 10 | ✗ | ✗ | ✗ | ✗ |
| PF-PASCAL | 20 | 1.3k | 4–17 | ✗ | ✗ | ✗ | ✗ |
| SPair-71k | 18 | 71k | 3–30 | ✗ | ✗ | ✗ | ✗ |
| DISCOBOX | 12 | 36k | 1–12 | ✗ | ✗ | ✓ | ✗ |
| MISC210K | 34 | 218k | 5–52 | ✗ | ✗ | ✗ | ✗ |
| **SOCO** | **100** | **1M** | **6–32** | ✓ | ✓ | ✓ | ✓ |

– **SOC as a representation diagnostic.** We conduct extensive experiments across a broad family of vision models, demonstrating that SOC correlates more strongly than ImageNet $k$NN with dense downstream tasks, positioning SOC as a practical zero-shot diagnostic of representation quality.

## 2 Related work

**Semantic Correspondence Benchmarks.** Finding correspondences is a fundamental task in computer vision, ranging from geometric [37,52] and stereo matching [44,55] to optical flow [11] and tracking [71], which are typically constrained to the same instance or scene. In contrast, semantic correspondence aims to establish correspondences between object parts *across* different instances of the same category. Early datasets such as PF-PASCAL and PF-WILLOW [27], TSS [65], and Freiburg-Cars [56] defined keypoint correspondences but they were limited in scale and category diversity. Zhang et al. [79] propose a semantic correspondence benchmark based on animal keypoints from AP-10K [77]. However, it does not include man-made objects, which have more diverse keypoint types and are equally important for probing general object-level understanding. SPair-71k [45] became the de-facto standard benchmark by providing 71k image pairs across 1,800 images from 10 rigid categories of PASCAL 3D+ [73] and 8 non-rigid categories of PASCAL VOC 2012 [21], out of which 481 images are used for testing. Due to the imbalanced class selection, quadruped animals and vehicles are favored. MISC210K [63] focuses on multi-instance correspondence and increases dataset scale, but its keypoints are defined by geometric heuristics rather than a hierarchical taxonomy of semantic concepts, which—as in SPair-71k—prevents cross-category evaluation. Additionally, current SC benchmarks do not provide keypoint descriptions, preventing systematic evaluation of LVLMs. SOCO addresses these limitations by introducing the concept of *Semantic Object Correspondence*, a taxonomy-driven formulation that specifically separates geometric from non-geometric semantic correspondences and standardizes what constitutes a valid correspondence across object categories. Based on this, we create a dataset of diverse categories with taxonomy-driven SC keypoints and textual descriptions, forming the basis for a more comprehensive benchmark.

**Semantic Correspondence in the Era of Foundation Models.** Self-supervised and multimodal foundation models have renewed interest in semantic correspondence as a probe for representation quality [20, 60, 66], after various studies have shown that features obtained from such models can be utilized for identifying semantic correspondences in a zero-shot manner [13, 25, 40, 47, 62, 64, 80], even though they do not encode the 3D part composition particularly well [14, 19, 42, 43, 61, 67, 79]. Evaluating semantic correspondence (SC) performance provides a complementary diagnostic to conventional tasks such as classification [16, 18, 22] or segmentation [15, 81]: by measuring how well models align object parts under appearance and pose variation, it reveals whether representations encode *fine-grained part-level* and *3D-aware* structure rather than local appearance details or global category cues.

In parallel to advances in SSL, vision–language models (VLMs) such as CLIP [48], BLIP [34], and Flamingo [1] were developed to align visual and textual modalities, but their evaluation focuses mainly on retrieval and captioning [48, 57] rather than fine-grained spatial understanding. Moreover, modern large vision-language models (LVLMs) such as LLava [38], Qwen-VL [5], GPT-4V [46], and Gemini [26], extend this paradigm toward multimodal visual reasoning, yet their evaluation remains dominated by high-level tasks like VQA [39, 78] and high-level spatial reasoning [75]. BLINK [24] contains a limited number of questions targeting semantic correspondence. However, since it is built on SPair-71k and does not contain language annotations, this benchmark does not provide a comprehensive evaluation of diverse fine-grained object understanding. Our work addresses this gap by introducing a benchmark that enables a systematic evaluation of LVLMs in terms of their visual correspondence and natural language alignment, allowing analysis of how linguistic cues influence fine-grained correspondence-level understanding.

## 3 A Taxonomy for Semantic Correspondence

Semantic correspondence (SC) is commonly understood as the task of matching points with similar semantics across different instances of an object category. However, the definition of "semantic" correspondence has remained vague and dataset-dependent. We detail this in Sec. 3.1. To address this gap, we propose a taxonomy for the SC task, providing a principled foundation for systematic annotation and evaluation. The proposed taxonomy forms the conceptual basis for **Semantic Object Correspondence (SOC)**, a formulation that explicitly separates the local semantics and geometric position of an object part. We introduce SOC in the following section (Sec. 3.2) and show how it resolves the inconsistencies observed in existing benchmarks.

### 3.1 Limitations of Current SC Keypoint Annotations

Existing SC object datasets (e.g., PF-PASCAL [27], MISC210K [63], Freiburg Cars [56], SPair-71k [45]) lack a systematic, hierarchical keypoint annotation

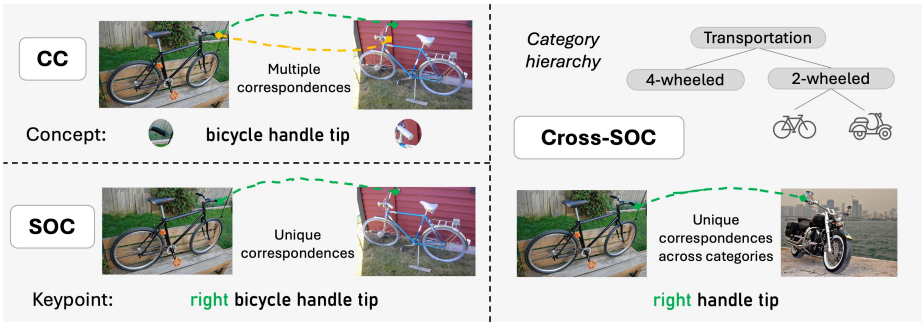



**Fig. 2: Illustration of concept correspondence (CC), semantic object correspondence (SOC), and cross-category semantic object correspondence (Cross-SOC).** SOCO differentiates CC and SOC, which define unique correspondences by disambiguating multiple instances of the same concept via geometric attributes, such as *right*. Cross-category matches (Cross-SOC) are derived from the accompanying category hierarchy.

strategy that scales across categories. Their annotations are often defined geometrically (e.g., midpoints on TV or boat contours) rather than as self-contained semantic concepts, are ambiguous for categories with large intra-class variability (*boats*) or symmetry (*bottles*, *potted plants*), are defined on 2D projections and thus break under viewpoint change, and are sometimes internally inconsistent (e.g., the "end" of a *train*). Crucially, current benchmarks evaluate object correspondence only *within* categories, ignoring relationships between semantically related objects (cars/trucks/buses) and thereby preventing assessment of cross-category semantic transfer. We illustrate concrete cases of annotation limitations in Fig. r3 in the supplementary.

These limitations are not merely annotation artifacts but stem from the absence of a structured representation of object parts. A principled formulation requires three properties: keypoints grounded in local semantics (unambiguous identification); *identity attributes* that distinguish repeated parts (front-left vs. rear-right wheel); and an explicit hierarchical organization of semantic concepts that can be reused across categories rather than redefined per class.

### 3.2 A Taxonomy of Semantic Object Correspondence

To introduce a more principled formulation of semantic correspondence, we define the term **Semantic Object Correspondence (SOC)**. SOC explicitly separates two complementary aspects: the local semantics of an object part and its spatial configuration within the overall object structure. This allows probing whether a model is able to match *semantic concepts* and *semantic object keypoints* that include a positional attribute.

A **semantic concept** is defined as a uniquely identifiable location (e.g., a corner point) within an object part that is typically shared across instances of the same category. Concepts capture the local semantics of a location on an object and its immediate functional context—for instance, the *door handle* of a car,

irrespective of whether it belongs to the left or right door. In contrast, **semantic object keypoints** are concrete, instance-specific realizations of a semantic concept. Each semantic object keypoint inherits from a concept but is further disambiguated by additional positional attributes that describe its placement within the object or component, such as *left*, *right*, *bottom*, or *rear*, which are consistently defined in the object-centric coordinate system. Concepts therefore describe *what* object part is being matched, whereas semantic object keypoints also specify *which instance* of that part within an object is considered. This makes finding correspondences among semantic object keypoints more challenging than concept-level matching, as a model must capture or reason about both semantic identity and geometric placement within the object context to correctly identify correspondences. While matching *concepts* across two instances (*concept correspondence, CC*) can yield non-unique matches across keypoints, keypoint matching (*semantic object correspondence, SOC*) always has a unique solution. Formally, a *Semantic Object Correspondence* is defined as a match between two keypoints that share the same semantic concept and identical object-relative identity attributes, ensuring both semantic and geometric matching.

Importantly, semantic concepts are not restricted to a single category (*cross-category SOC or Cross-SOC*). For example, a *wheel* concept may appear in a passenger car and in a school bus or tractor. To capture this hierarchical and cross-category structure, we propose to organize all semantic concepts within a taxonomy that spans categories, super-categories, and shared concepts among objects. This hierarchy enables correspondence evaluation both within categories and across related object classes. Fig. 2 illustrates CC, SOC, and Cross-SOC.

Together, this formulation establishes a coherent and extensible annotation framework for semantic correspondence, which forms the conceptual foundation for the **SOCO** dataset introduced next.

## 4 The SOCO dataset

Building on the taxonomy introduced in Sec. 3.2, we construct **SOCO**: a large-scale, taxonomy-driven dataset for evaluating *Semantic Object Correspondence (SOC)*. SOCO is designed to address key limitations of prior correspondence benchmarks by providing (1) a standardized, semantically grounded keypoint schema, (2) cross-category and hierarchical image-keypoint pairs, (3) and a substantially broader and more balanced set of object categories. Additionally, SOCO introduces *language descriptions* for all keypoints, enabling unified evaluation of both vision and vision–language correspondence models.

### 4.1 Dataset Creation

In the following, we describe the steps of the dataset creation: Image collection, category distribution, keypoint annotation, and language descriptions.

**Image collection.** All images are samples from ImageNet. We rely on 2D and 3D annotations from ImageNet3D [41] for man-made objects and on keypoint annotations from the Animal3D dataset [74] for the animal categories.

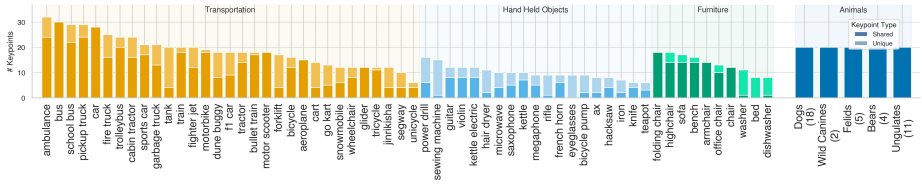



**Fig. 3: Statistics of labeled keypoints.** Keypoints in SOCO are annotated for a diverse set of categories from four super-categories. Each category is labeled with a subset of keypoints that are shared across multiple categories. The animal keypoints are shared across all animal categories.

We only retain images that (1) contain valid pose metadata, (2) depict a single salient object, and (3) have a sufficiently large object size.

**Category distribution.** SOCO comprises **100 categories** organized into four high-level super-categories: *Transportation* (31 classes), *Hand-held Objects* (20 classes), *Furniture* (9 classes), and *Animals* (40 classes).

**Keypoint annotation.** All keypoints follow the introduced taxonomy. While annotations for animal categories can be acquired from animal keypoint datasets, annotations of man-made objects that follow the taxonomy do not exist and, therefore, need to be collected. For this purpose, initial annotations are acquired via Amazon Mechanical Turk and refined through a manual verification stage. A user-friendly UI with integrated keypoint reference cards was developed to enable high-quality annotations. Three qualified annotators independently complete each image annotation, and the annotations are median-aggregated after removing outliers. Every keypoint annotation is verified manually to ensure consistency and accuracy. The median per-keypoint standard deviation across annotators is 0.85% (normalized by the maximum image dimension), indicating strong agreement. During manual verification, 65.4% of annotations required only minor refinements within PCK@0.05 tolerance, while 6.8% required larger corrections, e.g., due to confused conventions (e.g. left vs. right).

**Language Descriptions.** Each annotated keypoint includes a human-specified language description that combines its categorical, conceptual, and geometric attributes. Descriptions are generated programmatically using the tuple `(category, concept, keypoint position within the object part, object part position within the whole object)`, e.g., *"Center point of the front left wheel of a bus"*.

### 4.2 Dataset Statistics

Figure 3 presents the per-category keypoint distribution, including how many keypoints are shared with other categories. For each object category, 40 images are annotated, ensuring diverse viewpoints, shapes, and instance-level variations, resulting in a total of 4000 images. We construct Semantic Object Correspondence (SOC) pairs by matching images within the same category, requiring at least three shared semantic keypoints. This yields around 73k SOC pairs with a

**Table 2: Model performances on SOCO.** We report PCK@0.1 across concept correspondence (CC), semantic object correspondence (SOC), its cross-category variant (Cross-SOC), and the geometric variant (SOC-geo) as well as supercategory-wise results. As more geometric awareness is required for SOC and more semantic abstraction for Cross-SOC, model performance drops for all models. Model rankings change for geometric SOC, indicating that models encode object geometry and semantics to different extents. Additional evaluations are provided in the supplementary.

| Model | Accuracies by Tasks | | | | SOC by Supercategories | | | |
|---|---|---|---|---|---|---|---|---|
| | CC | SOC | Cross-SOC | SOC-geo | Trans. | Hand | Furn. | Animals |
| DINOv1 | 43.8 | 30.6$_{\downarrow 13.2}$ | 23.9$_{\downarrow 19.8}$ | 68.4 | 29.1 | 32.9 | 27.4 | 31.4 |
| DINOv2 | **78.9** | **60.4**$_{\downarrow 18.5}$ | **54.9**$_{\downarrow 24.0}$ | 71.9 | **56.9** | **61.6** | 45.5 | **66.3** |
| DINOv3 | 69.7 | 55.5$_{\downarrow 14.2}$ | 49.4$_{\downarrow 20.3}$ | 76.0 | 51.6 | 57.4 | **59.9** | 56.6 |
| iBOT | 55.2 | 39.6$_{\downarrow 15.5}$ | 34.1$_{\downarrow 21.1}$ | 69.4 | 36.1 | 40.1 | 32.8 | 43.9 |
| I-JEPA | 60.5 | 46.3$_{\downarrow 14.2}$ | 38.4$_{\downarrow 22.1}$ | 71.7 | 41.5 | 51.0 | 46.7 | 47.7 |
| C-RADIOv3 | 69.0 | 51.1$_{\downarrow 18.0}$ | 46.3$_{\downarrow 22.7}$ | 68.9 | 51.7 | 48.1 | 39.7 | 54.8 |
| DUNE | 60.1 | 45.7$_{\downarrow 14.4}$ | 39.4$_{\downarrow 20.7}$ | 73.5 | 40.0 | 50.5 | 51.0 | 46.6 |
| SD 2.1 | 60.5 | 49.0$_{\downarrow 11.5}$ | 42.7$_{\downarrow 17.8}$ | **76.8** | 46.4 | 47.0 | 50.9 | 50.9 |
| CroCov2 | 15.1 | 10.2$_{\downarrow 5.0}$ | 7.8$_{\downarrow 7.3}$ | 67.2 | 11.6 | 12.3 | 10.2 | 8.1 |
| MAE | 14.4 | 9.4$_{\downarrow 5.0}$ | 7.1$_{\downarrow 7.3}$ | 66.5 | 10.0 | 11.9 | 11.7 | 7.1 |
| PIXIO | 49.5 | 37.5$_{\downarrow 11.9}$ | 32.9$_{\downarrow 16.5}$ | 71.8 | 37.3 | 40.2 | 46.7 | 34.2 |
| CLIP | 25.2 | 16.2$_{\downarrow 9.0}$ | 11.4$_{\downarrow 13.8}$ | 64.7 | 17.9 | 14.7 | 11.7 | 16.8 |
| PE-Spatial | 63.2 | 45.9$_{\downarrow 17.4}$ | 40.9$_{\downarrow 22.4}$ | 68.5 | 46.5 | 43.4 | 38.7 | 48.4 |
| QWEN-L | 36.8 | 26.0$_{\downarrow 10.7}$ | 22.8$_{\downarrow 14.0}$ | 66.1 | 28.1 | 28.4 | 29.3 | 22.5 |

total of around 560k keypoint correspondences. Concept correspondences (CC) are generated using the same pairs.

We also form cross-category (Cross-SOC) pairs, using a minimum of three shared semantic keypoints. Due to the large combinatorial space of cross-category pairings, Cross-SOC generation results in around 1.3M cross-category correspondence pairs. These complementary pairing regimes (CC, SOC, and Cross-SOC) provide progressively more challenging correspondences that support evaluation across concepts, keypoints, and different categories.

## 5 Experiments

In this section, we benchmark several foundation models on semantic object correspondence. We first report results for vision encoders (Sec. 5.1) and LVLMs (Sec. 5.2). Then, we analyze how SOC relates to other vision tasks (Sec. 5.3).

### 5.1 Vision Foundation Model Evaluation on SOCO

**Evaluation Setup.** In the following, we evaluate common foundation models on SOCO and compare their performance on three subtasks: **First**, *concept correspondence* (CC) evaluates whether semantic concepts can be localized correctly. **Second**, *semantic object correspondence* (SOC) evaluates whether a model also encodes the geometric position of such a semantic concept relative to the whole object. **Third**, the most challenging cross-category setting Cross-SOC probes whether representations robustly encode the evaluated concepts across different

object categories. We evaluate on three fixed random SOCO subsets, which are released together with the full dataset. For each task (CC, SOC, Cross-SOC), we use 20k pairs with a uniform number of image pairs per category, ensuring high category and image diversity while keeping a manageable evaluation cost.

We select a representative set of current representation learning approaches: Self-supervised models like the DINO family [13,47,60], iBOT [82], I-JEPA [3], MAE [28], and PIXIO [76], vision models trained with text supervision [7,9,48] and with a multi-view reconstruction objective [70], a generative image diffusion model [53,64], and distilled models [29,50,54].

Following common practice in previous work [20,60,80], we evaluate SOC in a zero-shot manner: Given a source image $I^s$, a target image $I^t$, and a query point $p_i^s \in \mathbb{R}^2$ in the source image, the corresponding target point $p_i^t \in \mathbb{R}^2$ is computed by selecting the nearest feature vector in the target image through the argmax cosine similarity between the feature vector $f_i^s$ at the query point and the feature map $\mathcal{F}^{\mathrm{t}}$ of the target image:

$$p_i^t = \arg\max_{q_i^t \in I^t} \mathrm{sim}\big(f_i^s, \mathcal{F}^t(q_i^t)\big)\,. \tag{1}$$

For model evaluation, we follow common practice [45,79] and evaluate the matching performance via the Percentage of Correct Keypoints (PCK). It is defined by the ratio of correctly predicted keypoints that are within a radius of $R = \alpha \cdot \max(h, w)$ around the correct ground truth keypoint, where $h$ and $w$ refer to the height and width of the bounding box of the considered object, respectively. In the main paper, we report PCK at $\alpha = 0.1$, averaged over all image pairs of the dataset (*per-img*). We keep the image resolution fixed across models for fair comparison with small deviations if the specified image resolution is not a multiple of the given patch size. This benefits models with a smaller patch size. However, using a fixed number of patches does not substantially change model rankings, as we show in the supplementary.

As SOCO supports the explicit separation of geometric attributes and semantic concept, we further evaluate *SOC-geo*: This evaluates specifically whether a models is capable of differentiating the geometric positions of keypoints of the same concept. Given one source keypoint, the argmax is computed over all instances of the same concept for a target image of the same category. We only select image pairs where there are at least two pairs to match, which results in around $100k$ evaluated keypoint pairs. The random performance is 41.24% for this setting: The number of evaluated keypoints varies across categories and images. E.g., a car wheel might appear two or three times on an image but for a chair all four legs are typically visible.

**Experimental Results.** Model evaluation results are presented in Tab. 2 and we summarize the findings below.

Strong semantic representations in vision foundation models do not imply geometric part awareness.

This finding is indicated by the consistent and substantial performance drops from CC to SOC for all evaluated models. Notably, the magnitude of this drop

scales with overall model performance, suggesting that stronger semantic representations do not close the gap to geometric part awareness. This effect persists even for the best models (e.g., DINOv2): they capture semantic concepts well but struggle to disambiguate repeated object parts, as their representations do not reliably encode object-level geometry. Performance drops further in the cross-category setting Cross-SOC, as the appearance across labeled object parts changes even more strongly.

The per-supercategory results in Tab. 2 indicate substantially different performance across categories. Further, the CC→SOC gap varies: it is largest for *Furniture* (DINOv2: SOC 45.5 vs CC 77.5) and *Transportation*, where repeated symmetric parts, such as chair/table legs and the front/rear, left/right wheels of vehicles—dominate, and smaller for the more articulated but less repetitive *Animals* and the heterogeneous *Hand-held* super-categories. Interestingly, model rankings change with object structure: DINOv3 outperforms DINOv2 on *Furniture* (59.9 vs 45.5) despite being weaker on average, and SD 2.1 and DUNE also become comparatively stronger when repeated parts dominate.

The *SOC-geo* evaluations in Tab. 2 , explicitly separating geometric and semantic object part localization, indicate that ranking changes between different models. Similar to the previous observations for the furniture category, DINOv3 outperforms DINOv2 and SD2.1 achieves the best performance, which indicates that object part position is more effectively encoded here.

A single average score therefore hides *which* capability a model is missing—exactly the diagnostic value our taxonomy is designed to expose.

Dense self-supervised learning objectives lead to stronger semantic correspondence representations than global alignment objectives.

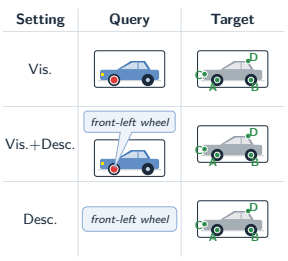
Representations from the DINO model family perform particularly well for concept correspondence (CC), indicating that their self-supervised objectives learn robust local semantic features. DINOv2 shows clear gains over DINOv1, whereas DINOv3 performs slightly worse across all correspondence settings. Models such as C-RADIOv3 and DUNE, which are distilled from strong dense feature encoders including DINOv2, inherit these properties and achieve competitive performance. In contrast, models trained with global alignment objectives, such as CLIP [48], perform substantially worse, reflecting the limited spatial precision of their representations. Compared with CLIP, the larger-scale PerceptionEncoder [9] improves correspondence performance in its spatial variant, consistent with the SOCO evaluation results. Interestingly, the vision encoder of Qwen2.5-VL [7] performs similarly poorly to CLIP on this task.

Finally, reconstruction-based models such as MAE and CroCoV2 perform poorly, as their objectives primarily encourage instance-specific appearance reconstruction rather than semantic feature alignment. However, PIXIO demonstrates that scaling reconstruction-based objectives can substantially improve dense correspondence representations. I-JEPA achieves comparatively strong performance despite being trained only on ImageNet-1k.

**Table 3:** SOCO evaluation results for LVLMs. All settings show the target image with candidate keypoints; only the query differs. *Vis.* uses a marked source image, *Vis.+Desc.* additionally provides the keypoint description, and *Desc.* uses only the keypoint description as query. Gray values denote the absolute difference to *Vis.*.

| Method | Vis. | Vis.+Desc. | Desc. |
|---|---|---|---|
| *Baselines* | | | |
| Random | 0.4 | 0.4 +0.0 | 0.4 +0.0 |
| Random++ | 25.0 | 25.0 +0.0 | 25.0 +0.0 |
| DINOv2 | 81.0 | 81.0 +0.0 | 81.0 +0.0 |
| *LVLMs* | | | |
| LLaVA-OV-7B [33] | 2.9 | 14.1 +11.2 | 24.3 +21.4 |
| InternVL3.5-8B [69] | 24.9 | 38.5 +13.6 | 39.6 +14.7 |
| Qwen2.5-VL-3B [7] | 5.2 | 17.4 +12.2 | 29.9 +24.7 |
| Qwen2.5-VL-7B [7] | 19.4 | 30.8 +11.4 | 39.1 +19.7 |
| Qwen3-VL-4B [6] | 8.6 | 18.0 +9.4 | 44.4 +35.8 |
| Qwen3-VL-8B [6] | 34.2 | 30.8 −3.4 | 54.0 +19.8 |
| GPT4o [30] | 30.2 | 30.9 +0.7 | 37.6 +7.4 |

LVLM evaluation settings.




### 5.2 LVLM evaluation on SOC

In this section, we analyze several representative LVLMs on SOC and compare their performance in settings with and without access to textual descriptions.

**Experimental Setup.** Following the BLINK benchmark [24], we formulate semantic correspondence as a multiple-choice VQA task. We adopt the *CircularEval* protocol [39], where each question is presented to the LVLM four times with different permutations of the answer choices (ABCD) to enforce a consistent prediction. An answer is considered correct only if the model predicts the correct option in all permutations, and we report accuracy under this strict criterion.

Unlike BLINK, which uses only an annotated reference image as the visual prompt, we further study how different query prompt types affect semantic correspondence performance. Specifically, we evaluate three settings: *Vis.*, *Vis.+Desc.*, and *Desc.* In all three settings, the target image with candidate keypoint markers A/B/C/D is shown to the LVLM; the settings differ only in how the query keypoint is specified (*cf*. the inset of Tab. 3). In *Vis.*, the query is provided as a visual prompt, where the query keypoint is marked with visual markers in the source image. In *Vis.+Desc.*, a textual description of the query keypoint is provided additionally. In *Desc.*, the source image is omitted, and the query keypoint is specified using only the textual description. The target image and its A/B/C/D candidate markers remain visible. This results in a *Random++* baseline performance of 25%, where the predictions are consistently permuted. The gap between *Random++* and *Vis.* therefore quantifies cross-image visual matching, while *Desc.* measures text-prompted keypoint localization in the target image. Full prompts and additional illustrations are provided in the supplementary. We evaluate the LVLMs on a smaller subset of SOCO with 20 image pairs per category, and adapt DINOv2 to the same 4-choice protocol by selecting the candidate patch with the highest cosine similarity to the query feature. The quantitative results on SOCO are summarized in Tab. 3. As the evaluation

follows a circular protocol, the *Random++* baseline returns a random answer that is consistent across the four permuted questions of the same evaluation.

**Experimental Results.** LVLM evaluation results are presented in Tab. 3, and findings are discussed below.

LVLMs are stronger at text-prompted keypoint localization than at visual-reference cross-image matching, exposing a gap between language-grounded localization and fine-grained visual correspondence.

A consistent trend across LVLMs is that providing an explicit keypoint description (*Vis.+Desc.* and *Desc.*) improves performance compared to a purely visual query (*Vis.*). Notably, all models achieve higher accuracy in the description-only setting (*Desc.*) than in the visual-reference setting (*Vis.*). This indicates that LVLMs are more effective at localizing a textually described part within a single image than at transferring a marker from a source image to the target image.

Overall, recent models show clear improvements in both visual and language understanding. For example, the Qwen family shows consistent gains from smaller to larger models, and the Qwen3-VL-8B model outperforms its Qwen2.5-VL-7B predecessor, indicating that scaling and improved training pipelines translate into stronger semantic correspondence capabilities.

However, DINOv2 adapted to the 4-choice setting outperforms all evaluated LVLMs, achieving 81.0% compared to the best evaluated LVLM at 54.0%. This suggests that current LVLMs rely heavily on textual guidance but remain limited in their ability to align visual and textual modalities for fine-grained, cross-image correspondences. Therefore, despite recent progress, semantic correspondence on SOCO remains a challenging task for current LVLMs.

### 5.3 Relation to Other Vision Downstream Tasks

The previous sections evaluated SOC across a diverse set of models. In contrast, vision foundation models are typically assessed on various downstream tasks [9,47,50,60,66], spanning global objectives (e.g., image classification) and dense prediction tasks (e.g., tracking and semantic segmentation). These tasks require different evaluation protocols, such as linear probing for ImageNet [47], task-specific fine-tuning, or DPT-based training [49,60], and their outcomes can depend strongly on hyperparameter choices.

Currently, ImageNet still remains the gold standard task for measuring representation quality [3,47,60], as it correlates well to other tasks [32]. However, Bolya et al. [9] have shown that capturing global representations is not necessarily aligned with strong dense semantic features.

As SOC probes dense semantic and geometric features, it is more indicative of structured visual understanding than classification-based metrics such as ImageNet kNN, while remaining practical through a simple zero-shot protocol without hyperparameter tuning. We therefore study its relation to other semantic and geometric vision tasks to assess whether it can serve as a representative diagnostic of representation quality.

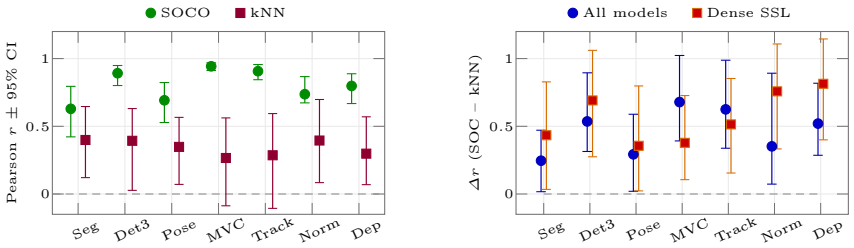



**Fig. 4: Per-task Pearson $r$ across 37 vision models, with 95% bootstrap CIs.** *Left:* SOC correlates with every downstream task more strongly than ImageNet kNN. *Right:* the SOC advantage $\Delta r = r_{\text{SOC}} - r_{\text{kNN}}$ stays positive on all tasks and is preserved on a 17 subset only including models trained with dense SSL objectives.

**Experimental Setup.** We evaluate the representational quality of modern vision and vision–language backbones on a representative set of tasks using a unified experimental protocol that builds directly on Probe3D [20]. We extend Probe3D with additional probes, including semantic object correspondence on SOCO, semantic segmentation [81], tracking [17], 3D pose estimation [41], and 3D object detection using an adapted version of the Omni3D [10] pipeline. Furthermore, we integrate a diverse set of vision foundation models, enabling performance evaluation at large scale. This unified design allows us to evaluate both fine-grained and object-level 3D understanding under identical backbone, decoding, and optimization conditions. We evaluate *depth estimation* and *surface normal prediction* on NYU [59], *geometric multi-view correspondence* on NAVI [31], k-nearest neighbor *kNN classification* on ImageNet [16], *3D pose regression* on ImageNet3D [41], *semantic segmentation* on ADE-20k [81], and *zero-shot tracking* on TAP-Vid [17], covering a wide spectrum of monocular single- and multi-view spatial reasoning requiring semantic and/or geometric understanding. We largely follow the hyperparameters used by El Banani et al. [20] and discuss implementation details in the supplementary. We open-source the evaluation framework.

**Experimental Results.** We compute the Pearson correlation between SOC performance and the downstream metrics across 37 vision models, with 95% bootstrap CIs (10k resamples) and leave-one-out checks. The results are summarized in Fig. 4.

SOC has a stronger correlation to various dense geometric and semantic tasks than kNN ImageNet classification.

SOC dominates kNN on every evaluated downstream task (Fig. 4), with CIs that exclude zero for the six conclusive metrics. The advantage of SOC over kNN persists after restricting the pool to 17 dense-SSL models, ruling out a dense-vs-global confound. Leave-one-out resampling agrees with the full-pool results on every metric. Overall, this suggests that SOC is a practical zero-shot diagnostic that is more aligned with dense vision tasks than ImageNet kNN.

## 6 Conclusion

We introduced Semantic Object Correspondence (SOC), a principled formulation of semantic correspondence that explicitly models the relationship between object parts and the overall object structure, providing a clearer separation between geometric matching and semantic object-level understanding. Building on this formulation, we developed SOCO, a large-scale benchmark that provides hierarchical part annotations, cross-category correspondences, and accompanying language descriptions, thus addressing the core limitations of existing datasets.

Through extensive evaluation of vision and multimodal foundation models, we demonstrated that SOCO exposes differences in their ability to capture fine-grained, object-centric structure. The taxonomy makes three failure modes separately measurable: the CC→SOC gap isolates repeated-part disambiguation, the SOC→Cross-SOC gap isolates category-specific concept encoding, and the Vis. vs. Desc. gap in LVLMs separates cross-image visual matching from language-grounded part localization. Our results show that: (1) models reliably match semantic concepts but struggle with object-level geometry; (2) cross-category correspondence remains challenging even for the strongest vision backbones; (3) large vision–language models are stronger at text-prompted keypoint localization than at visual-reference correspondence, revealing a gap between language-grounded localization and visual matching; and (4) SOC performance correlates with dense vision tasks more strongly than ImageNet $k$NN, making SOC a powerful zero-shot diagnostic for representation quality.

SOCO provides a unified testbed for analyzing structured part-level visual and multimodal understanding in modern foundation models. We hope it serves as a stepping stone toward models that not only recognize objects but also understand their parts and structural relationships in a way that generalizes across categories and modalities.

## Acknowledgments

AK acknowledges support via his Emmy Noether Research Group funded by the German Research Foundation (DFG) under grant number 468670075. We thank Matthis Heimberg for early analyses and experiments.

## References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millicah, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: A visual language model for few-shot learning. NeurIPS **35** (2022)
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: CVPR. pp. 3686–3693 (2014)

3. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: CVPR. pp. 15619–15629 (2023)
4. Aydemir, G., Xie, W., Güney, F.: Can visual foundation models achieve long-term point tracking? In: ECCV Workshops (2024)
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
6. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
7. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
8. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., Shulman, E.: ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. arXiv preprint arXiv:2111.08897 (2021)
9. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. arXiv preprint arXiv:2504.13181 (2025)
10. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: Omni3D: A large benchmark and model for 3D object detection in the wild. In: CVPR (2023)
11. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV. pp. 611–625 (2012)
12. Cai, Z., Yeh, C.F., Xu, H., Liu, Z., Meyer, G., Lei, X., Zhao, C., Li, S.W., Chandra, V., Shi, Y.: DepthLM: Metric depth from vision language models. arXiv preprint arXiv:2509.25413 (2025)
13. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
14. Chi, Y., Sommer, L., Dünkel, O., Muhle, D., Cremers, D., Theobalt, C., Kortylewski, A.: C3PO: Canonicalization of 3D pose from partial views with generalizable correspondence features. In: 3DV (2026)
15. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
17. Doersch, C., Gupta, A., Markeeva, L., Recasens, A., Smaira, L., Aytar, Y., Carreira, J., Zisserman, A., Yang, Y.: TAP-Vid: A benchmark for tracking any point in a video. NeurIPS **35**, 13610–13626 (2022)
18. Dünkel, O., Jesslen, A., Xie, J., Theobalt, C., Rupprecht, C., Kortylewski, A.: CNS-Bench: Benchmarking image classifier robustness under continuous nuisance shifts. In: ICCV (2025)
19. Dünkel, O., Wimmer, T., Theobalt, C., Rupprecht, C., Kortylewski, A.: Do it yourself: Learning semantic correspondence from pseudo-labels. In: ICCV (2025)
20. El Banani, M., Raj, A., Maninis, K.K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., Jampani, V.: Probing the 3D awareness of visual foundation models. In: CVPR. pp. 21795–21806 (2024)

21. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes challenge: A retrospective. IJCV **111**(1), 98–136 (2015)
22. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. IJCV **88**(2), 303–338 (2010)
23. Florence, P.R., Manuelli, L., Tedrake, R.: Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. arXiv preprint arXiv:1806.08756 (2018)
24. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: BLINK: Multimodal large language models can see but not perceive. In: ECCV (2024)
25. Gan, C., Tu, Y., Chen, X., Chen, T., Li, Y., Harandi, M., Lin, W.: Unleashing diffusion transformers for visual correspondence by modulating massive activations. In: NeurIPS (2025)
26. Gemini Team: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
27. Ham, B., Cho, M., Schmid, C., Ponce, J.: Proposal flow. In: CVPR. pp. 3475–3484 (2016)
28. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
29. Heinrich, G., Ranzinger, M., Yin, H., Lu, Y., Kautz, J., Tao, A., Catanzaro, B., Molchanov, P.: RADIOv2.5: Improved baselines for agglomerative vision foundation models. In: CVPR (2025)
30. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A.J., Welihinda, A., Hayes, A., Radford, A., et al.: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
31. Jampani, V., Maninis, K.K., Engelhardt, A., Karpur, A., Truong, K., Sargent, K., Popov, S., Araujo, A., Martin-Brualla, R., Patel, K., Vlasic, D., Ferrari, V., Makadia, A., Liu, C., Li, Y., Zhou, H.: NAVI: Category-agnostic image collections with high-quality 3D shape and pose annotations. In: NeurIPS (2023)
32. Kornblith, S., Shlens, J., Le, Q.V.: Do better ImageNet models transfer better? In: CVPR. pp. 2661–2671 (2019)
33. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. TMLR (2025)
34. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
35. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 936–944 (2017)
36. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755 (2014)
37. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: SIFT flow: Dense correspondence across different scenes. In: ECCV. pp. 28–42 (2008)
38. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
39. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: MMBench: Is your multi-modal model an all-around player? In: ECCV. pp. 216–233 (2024)
40. Luo, G., Dunlap, L., Park, D.H., Holynski, A., Darrell, T.: Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In: NeurIPS (2023)

41. Ma, W., Zhang, G., Liu, Q., Zeng, G., Kortylewski, A., Liu, Y., Yuille, A.: ImageNet3D: Towards general-purpose object-level 3D understanding. NeurIPS **37**, 96127–96149 (2024)
42. Mariotti, O., Du, Z., Bhalgat, Y., Mac Aodha, O., Bilen, H.: Jamais vu: Exposing the generalization gap in supervised semantic correspondence. arXiv preprint arXiv:2506.08220 (2025)
43. Mariotti, O., Mac Aodha, O., Bilen, H.: Improving semantic correspondence with viewpoint-guided spherical maps. In: CVPR. pp. 19521–19530 (2024)
44. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: CVPR. pp. 4040–4048 (2016)
45. Min, J., Lee, J., Ponce, J., Cho, M.: SPair-71k: A large-scale benchmark for semantic correspondence. arXiv preprint arXiv:1908.10543 (2019)
46. OpenAI, Applin, S., Adesso, G., Ashfaq, R., Bai, M., Brammer, M., Fecht, E., Goodman, A., Grossman, S., Groh, M., Kirk, H.R., Gunitsky, S., Huang, Y., Kahn, L., Kumar, S., Madrid-Morales, D., Motoki, F., Ovadya, A., Peters, U., Robinson, M., Röttger, P., Wasserman, H., Wehsener, A., Walker, L., Vidgen, B., Zhu, J.: GPT-4V(ision) system card. Tech. rep., OpenAI (2023)
47. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. TMLR (2024)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
49. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021)
50. Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: AM-RADIO: Agglomerative vision foundation model reduce all domains into one. In: CVPR. pp. 12490–12500 (2024)
51. Ranzinger, M., Heinrich, G., McCarthy, C., Kautz, J., Tao, A., Catanzaro, B., Molchanov, P.: C-RADIOv4 technical report. arXiv preprint arXiv:2601.17237 (2026)
52. Rocco, I., Arandjelovic, R., Sivic, J.: Convolutional neural network architecture for geometric matching. In: CVPR. pp. 6148–6157 (2017)
53. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
54. Sarıyıldız, M.B., Weinzaepfel, P., Lucas, T., De Jorge, P., Larlus, D., Kalantidis, Y.: DUNE: Distilling a universal encoder from heterogeneous 2D and 3D teachers. In: CVPR. pp. 30084–30094 (2025)
55. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV **47**(1), 7–42 (2002)
56. Sedaghat, N., Brox, T.: Unsupervised generation of a viewpoint annotated car dataset from videos. In: ICCV (2015)
57. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K.: How much can CLIP benefit vision-and-language tasks? In: ICLR (2022)
58. Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does CLIP know about a red circle? visual prompt engineering for VLMs. In: ICCV (2023)

59. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGB-D images. In: ECCV (2012)
60. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. arXiv preprint arXiv:2508.10104 (2025)
61. Sommer, L., Dünkel, O., Theobalt, C., Kortylewski, A.: Common3D: Self-supervised learning of 3D morphable models for common objects in neural feature space. In: CVPR. pp. 6468–6479 (2025)
62. Stracke, N., Baumann, S.A., Bauer, K., Fundel, F., Ommer, B.: CleanDIFT: Diffusion features without noise. In: CVPR. pp. 117–127 (2025)
63. Sun, Y., Huang, Y., Guo, H., Zhao, Y., Wu, R., Yu, Y., Ge, W., Zhang, W.: MISC210K: A large-scale dataset for multi-instance semantic correspondence. In: CVPR. pp. 7121–7130 (2023)
64. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. NeurIPS **36**, 1363–1389 (2023)
65. Taniai, T., Sinha, S.N., Sato, Y.: Joint recovery of dense correspondence and cosegmentation in two images. In: CVPR. pp. 4246–4255 (2016)
66. Venkataramanan, S., Pariza, V., Salehi, M., Knobel, L., Gidaris, S., Ramzi, E., Bursuc, A., Asano, Y.M.: Franca: Nested matryoshka clustering for scalable visual representation learning. arXiv preprint arXiv:2507.14137 (2025)
67. Wandel, K., Wang, H.: SemAlign3D: Semantic correspondence between RGB images through aligning 3D object-class representations. In: CVPR (2025)
68. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6D object pose and size estimation. In: CVPR. pp. 2642–2651 (2019)
69. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
70. Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., Revaud, J.: CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow. In: ICCV (2023)
71. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: CVPR. pp. 2411–2418 (2013)
72. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2 (2019), software
73. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond PASCAL: A benchmark for 3D object detection in the wild. In: WACV. pp. 75–82 (2014)
74. Xu, J., Zhang, Y., Peng, J., Ma, W., Jesslen, A., Ji, P., Hu, Q., Zhang, J., Liu, Q., Wang, J., et al.: Animal3D: A comprehensive dataset of 3D animal pose and shape. In: ICCV. pp. 9099–9109 (2023)
75. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember and recall spaces. arXiv preprint arXiv:2412.14171 (2024)
76. Yang, L., Li, S.W., Li, Y., Lei, X., Wang, D., Mohamed, A., Zhao, H., Xu, H.: In pursuit of pixel supervision for visual pre-training. arXiv preprint arXiv:2512.15715 (2025)
77. Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D.: AP-10K: A benchmark for animal pose estimation in the wild. arXiv preprint arXiv:2108.12617 (2021)

78. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In: CVPR (2024)
79. Zhang, J., Herrmann, C., Hur, J., Chen, E., Jampani, V., Sun, D., Yang, M.H.: Telling left from right: Identifying geometry-aware semantic correspondence. In: CVPR. pp. 3076–3085 (2024)
80. Zhang, J., Herrmann, C., Hur, J., Polania Cabrera, L., Jampani, V., Sun, D., Yang, M.H.: A tale of two features: Stable Diffusion complements DINO for zero-shot semantic correspondence. NeurIPS **36**, 45533–45547 (2023)
81. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: CVPR. pp. 633–641 (2017)
82. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: iBOT: Image BERT pre-training with online tokenizer. In: ICLR (2022)