WHAT MATTERS FOR PERCEIVING OBJECT-PARTS IN VISION ENCODERS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent works provide evidence that large scale pretraining makes vision and language models organize data similarly, however this similarity is exclusively studied on scene- and object-level. This leaves the unanswered question: *are object-parts organized in vision models similar to language models?* Here, we focus on parts, as user-defined components of 3D objects. Our work tries to answer this question through a series of analyses:

(i) Part-similarity. We measure part-level representation similarity between frozen vision and language models using a modality agnostic protocol. To this end, we conduct the first large scale comparison of 39 vision encoders across 3 different vision modalities. **(ii) Part-alignment.** We further propose to test the ease of transforming vision models to language in the absence of parallel part annotations. To this end, we propose a simple probe based on differentiable optimal transport, particularly suited for data-scarce settings like 3D where large-scale paired part annotations are not avaiable. **(iii) Part-transferability.** Finally, we find that our probe can transfer to unseen part hierarchies when trained on bags of discovered part masks and cheaply obtained VLM-provided part names.

Our analyses demonstrate how visual pretraining strategy, architectural inductive biases, and modality influence both similarity and alignment with language on a part level with implications towards real-world applications.

1 INTRODUCTION

Existing analyses that investigate similarity between vision and language models study this at the level of complete scenes or objects. They compare representations across images (Huh et al., 2024; Gröger et al., 2026; Koepke et al., 2026; Schnaus et al., 2025) and point clouds (Hadgi et al., 2025) to paired language captions using global structural similarity metrics. These studies provide valuable evidence of coarse vision-language similarity; nonetheless, it remains unclear whether such similarity extends to the object-part level, a question with direct implications for tasks like fine-grained robotic control, content generation, and editing.

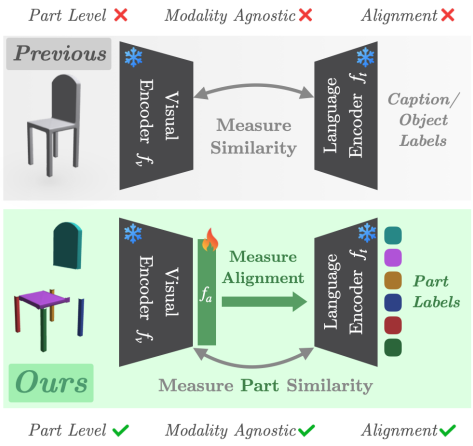


Figure 1: Previous studies have focused primarily on complete scene/object representations and remain limited in the amount of tested modalities and probing methods. We address this gap with the first comprehensive part-similarity and part-alignment analysis across 3 visual modalities.

However, there currently exists no diagnostic tool or systematic benchmark for evaluating part-level representation similarity across a variety of visual encoders. Our study fills this gap by asking: *Do visual representations exhibit **similarity** with language models at a **part** level?* To this end, we address part-similarity with established similarity metrics (Kornblith et al., 2019) under a unified modality agnostic protocol. More specifically, our protocol uses PartNet as the common comparison ground where provided objects have been annotated by humans

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

with associated part labels and masks. Using these labels and parsed geometries, our study uncovers how factors like visual modality, pretraining strategy, multi-modal pretraining, and architectural inductive biases affect part-similarity.

While part-similarity quantifies representational similarity between vision and language, modern vision-language systems rarely operate on pretrained representations directly. Thus, we additionally probe the ***part-alignment*** of these representations, which we define as the ease with which representations from one modality can be transformed into another. In data-scarce regimes, learning such alignment with supervision is challenging due to lack of paired data, thus our proposed probe is based on differentiable optimal transport, which can be optimized without part-level correspondences.

This unpaired setting is especially relevant when part annotations are scarce such as in 3D. Since our probe is general and requires no part-level correspondences we test whether it can transfer to unseen part hierarchies in our ***part-transferability*** study. To this end, we train our probe on *bags* of discovered part masks and VLM-provided part names. Our simple experiment shows strong transferability suggesting that unpaired matching may be a viable route for scaling part-labeling pipelines in data-scarce regimes.

Our contributions can be summarized as follows:

- We develop a modality-agnostic protocol with which we conduct the first comprehensive part-similarity and part-alignment analysis between 39 different vision models across 3 modalities.
- We make two major findings: (i) Part-similarity with language can emerge without multimodal pretraining and (ii) Strong part-similarity does not imply high part-alignment.
- We find that the unpaired probe transfers to unseen part hierarchies when trained on independently discovered part masks and VLM-provided part names.

2 Related Work

Representational Similarity. Emergent representational similarity across deep learning modalities has recently attracted substantial interest. The platonic representation hypothesis (Huh et al., 2024) posits that representations converge across domains, measuring pairwise distances between datapoints in vision and language encoders. Follow-up work examines convergence at varying scales (Koepke et al., 2026), identifies confounders such as model depth and width (Gröger et al., 2026), and analyzes misalignment in multimodal language models (Shu et al., 2025). Related applications leverage pairwise similarities for unsupervised image-caption matching via quadratic assignment (Schnaus et al., 2025), and extend the analysis to emergent alignment between complete 3D objects and language (Hadgi et al., 2025). However, none of the work has studied representational similarity at a fine-grained *part-level*, and none has compared multiple visual modalities with the same probe.

Fine-grained Analysis of Vision-Language Models. Recent work shows that multimodal large language models still struggle with fine-grained visual understanding (Tong et al., 2024), visual grounding (Peng et al., 2023), and faithful perception (Bai et al., 2025). Closer to our setting, Jiang et al. (2025) study patch-level alignment in multimodal LLMs, but restrict analysis to models with explicitly aligned vision and language components and extract aligned patches using pseudo masks and vision-language correspondences. This leaves open whether part-level similarity can emerge between independently trained vision and language models, and to what extent such correspondence is inherent to the learned representations rather than a through multimodal pretraining strategies.

3D and Language. Supervised 3D part segmentation typically assumes paired part segments and labels against expert-defined hierarchies (Qi et al., 2017; Mo et al., 2019; Wu et al., 2024), which remain scarce and costly to obtain. Recent work reduces this burden with cross-modal pipelines that lift 2D foundation-model segments and obtain pseudo labels from large vision-language models (Ma et al., 2024; Garosi et al., 2025; Yang et al., 2024), or that rely on prior data engines and limited human part datasets (Paul et al., 2026; Hadgi et al., 2026). These approaches still depend on paired segments and part labels during training or label generation. In contrast, our part-transferability

study asks whether part correspondence can be learned from unpaired segments and part names obtained independently.

3 PRELIMINARIES

Let $\mathcal{D} = \{(x_i, Y_i, \mathcal{T}_i)\}_{i=1}^{N}$ denote a dataset of visual observations, where $x_i \in \mathcal{X}$ may correspond to an image, video, or 3D object. Each observation is annotated with a set of semantic parts where

$$Y_{i,k} \subset \Omega_i$$

denotes the spatial support (or simply the annotated mask) of the k-th part over the observation domain Ω_i. Associated with each part is a semantic concept

$$\mathcal{T}_i = \{t_{i,1}, \ldots, t_{i,K_i}\},$$

where $t_{i,k}$ denotes the language label corresponding to part $Y_{i,k}$ for $k \in \{1, \ldots, K_i\}$. Given the visual encoder f_v we can define the dense visual features per datapoint as:

$$f_v(x_i) = F_i,$$

where $F_i \in \mathbb{R}^{|\Omega_i| \times d}$. While this feature provides global representation for the data sample (e.g. image or point cloud of a Chair). Part-level analysis requires pooling of features belonging to a single semantic concept (e.g. Leg, Armrest) defined as:

$$\mathbf{z}_{i,k} = \mathrm{Pool}\left(F_i[Y_{i,k}]\right),$$

where $F_i[Y_{i,k}]$ denotes the subset of features belonging to the k-th region and $\mathrm{Pool}(\cdot)$ is an aggregation operator (e.g., average pooling). Language embeddings per part feature are simply extracted from the langugage model f_l.

$$\mathbf{e}_{i,k} = f_l(t_{i,k}).$$

Hence, the latent representations that we compare throughout the rest of our study become the following:

$$\{\mathbf{z}_{i,k}\}_{k=1}^{K_i} \qquad \text{and} \qquad \{\mathbf{e}_{i,k}\}_{k=1}^{K_i}.$$

For brevity we skip the sample index i from the rest of this section.

3.1 ISOMORPHIC REPRESENTATION SIMILARITY

A common approach to comparing independently trained representation spaces is through *isomorphic representation metrics*. Rather than directly comparing embeddings, these methods compare the intrinsic geometry of each representation space. Recent works evaluating employ global metrics such as Centered Kernel Alignment (`CKA`), Singular Vector Canonical Correlation Analysis (`SVCCA`), or local metrics such as Mutual k-Nearest Neighbors (`Mutual-kNN`) to quantify this structural correspondence. Intuitively, `CKA` measures how similarly the encoded spaces structure pairs of data points; this computation requires known paired correspondences. We defer the formal definition to the supplementary (Section C).

3.2 OPTIMAL TRANSPORT

When part-level correspondences are unknown, pairwise metrics such as `CKA` no longer apply. Optimal Transport (OT) instead casts matching as moving probability mass from one set of features onto another at minimum cost: each visual part embedding is soft-assigned to language concepts through a transport plan. In our setting, a lower transport cost after a learned mapping indicates higher part-alignment. To optimize such mappings end-to-end, we use entropy-regularized (differentiable) OT, solved via Sinkhorn iterations (Cuturi, 2013). We defer the formal definitions of OT and differentiable OT to the supplementary (Section D).

4 MODALITY AGNOSTIC BENCHMARK DESIGN

In this section, we formalize the design of our benchmark and protocols that allow us to compare similarity and alignment across different modalities of vision encoders. We first motivate our choice of comparison domain and dataset in Section 4.1, describe the feature extraction process in Section 4.2, and summarize the vision and language models used in Section 4.3 and Section 4.4.

4.1 COMPARISON DOMAIN AND DATASET

Given that the modalities operate in different visual domains (2D, 3D, and 2D video), a unified analysis requires lifting model features into a common representation. We adopt the 3D domain for this purpose, from which static 2D images or orbiting video frames can be faithfully projected, and use point clouds as the shared grounds for comparison. This choice motivates PartNet (Mo et al., 2019) as our evaluation dataset: it provides large-scale 3D objects with hierarchical part segmentations, so geometry and language concepts can be associated at the part level under a consistent annotation scheme. Our experiments consider 25,900 objects across 24 categories with a total of 256,675 parts. PartNet's comprehensive and consistent part annotations make it especially suitable for probing semantic similarity between vision and language, as it provides a canonical setting for associating fine-grained geometry (parts) with language concepts. For all testing we use the official testing split for PartNet composed of 20% of the dataset. In sections with a learnable probe the rest 80% is used for training.

4.2 MODALITY-AGNOSTIC FEATURE EXTRACTION PROTOCOL

To enable a consistent comparison of representations across 3D, image, and video encoders, we first express every visual observation as a set of features associated with the same underlying 3D point cloud. Specifically, given an a mesh from PartNet x_i, we define a set of surface points $P_i = \left\{p_i^{(j)}\right\}_{j=1}^{N_i}$, where each point corresponds to a spatial location in the object and is associated with a single semantic label through the annotation masks Y_i introduced in Section 3. This produces a modality-independent feature representation

$$F_i \in \mathbb{R}^{N_i \times d},$$

where the j-th row of F contains the feature vector assigned to point $p_i^{(j)}$. Consequently, the notation introduced in Section 3 remains unchanged regardless of the underlying visual modality, and the subsequent part-level pooling is always performed over subsets $F_i[Y_{i,k}]$ (i.e. points associated with part k).

3D encoders already produce per-point features on P_i, whereas image and video encoders require an additional lifting step that back-projects dense feature maps onto the same point cloud given known camera poses. Within this, we use the multi-view lifting procedure of Garosi et al. (2025) to lift 2D features onto the 3D point cloud, a common strategy in 3D tasks that leverage 2D model priors. For simplicity, we choose average pooling as the $\mathrm{Pool}(\cdot)$ operator. We describe the lifting procedure for each modality in additional details (e.g., runtime, renderering procedure, view sampling, model choices etc.) in the Section A. We also find that our multi-view protocol was robust to pooling, in that using a single-view to measure similarity ranked vision encoders similarly (Section F).

4.3 VISION MODELS

We evaluate a diverse suite of frozen vision encoders spanning image, point-cloud, and video modalities, covering self-supervised, reconstruction-based, and vision–language (VL) pretrained models. Unless otherwise noted, we use publicly released checkpoints and extract dense per-point features under the protocol in Section 4.2. The complete model list and citations are deferred to Table 2 so that the main text can refer to models by name.

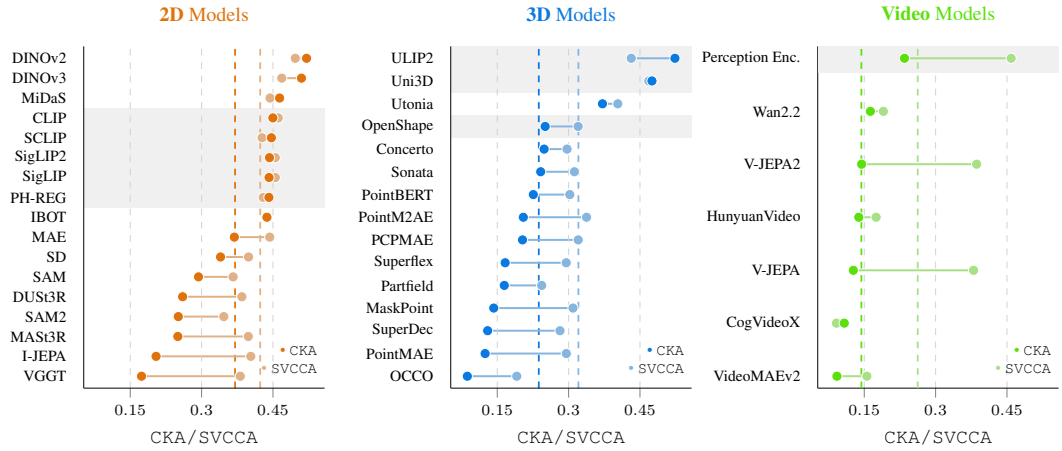


Figure 2: Part-Similarity using CKA/SVCCA scores for different visual models, grouped by modality. Shaded rows indicate models with explicit multi-modal pretraining.

4.4 Language Models

We follow Paul et al. (2026) and extract part-name embeddings ($\mathbf{e}_{i,k}$) from a sentence embedding transformer (all-mpnet-base-v2 (Song et al., 2020)). We fix this across majority of our experiments however provide a comprehensive ranking across other language models in a proceeding section.

5 Part-Similarity

Prior analyses of vision-language similarity focus on complete scenes or objects, however, it remains an open question whether visual representations are similar to language at the level of object parts. We therefore quantify part-similarity under our modality-agnostic protocol with two complementary metrics: CKA, which measures global representational similarity across part embeddings, and SVCCA (10 components), which compares the dominant shared subspaces after SVD. As a complementary view of local neighborhood structure, we also report Mutual-kNN ($k = 10$) in the supplementary (Section E); further implementation details follow Huh et al. (2024) unless stated otherwise.

5.1 Part-Similarity across vision models and modalities

In Figure 2 we report the CKA and SVCCA scores for different models, grouped by modality. The two metrics produce similar model rankings (Spearman $\rho = 0.90$); we therefore use the CKA to rank vision encoders in the following analysis.

3D encoders. 3D encoders can be geometrically part-aware, yet part-similarity with language varies widely across pretraining regimes. CLIP-inspired models such as ULIP-2 and Uni3D (gray rows) rank highest. OpenShape follows a similar pretraining recipe yet trails by a wide margin. We attribute this gap to its more modest encoder size and reliance on object-level captions, which may be less suited to part-level semantic structure. Recent point transformers that mix visual domains and modalities during training form a second tier; Utonia is notable here for high similarity from point-cloud-only training across object-centric, indoor, and outdoor 3D data, without language or 2D supervision. Conversely, masked reconstruction models (Point-MAE, Point-BERT, PCP-MAE) fall well below average under more limited object-level training data, underscoring that pretraining scale affects part-level similarity to language, not only object-level performance. Methods focused on segmentation or geometric decomposition (Partfield, SuperDec, SuperFlex) also lag, despite objectives that target part structure more directly.

2D encoders. Self-supervised vision-only transformers trained with patch-level objectives: DINOv2 and DINOv3 lead the 2D ranking and fall just short of ULIP-2. Despite explicit language

alignment, CLIP-style vision-language models lag these self-supervised counterparts, showing that object- or scene-level contrastive pretraining may not always be sufficient for part-level similarity. CLIP successors aimed at more refined dense features (SCLIP; PH-REG, SigLIP; SigLIP2) do not yield a clear improvement, though their ranks remain competitive. Similar to the 3D results, 2D segmentation encoders (SAM, SAM 2) rank near the bottom. This suggests that features that distinguish object parts need not organize them similarly to language. Feed-forward reconstruction models (MASt3R, DUSt3R, VGGT) also score poorly, except for MiDaS.

Video encoders. Under the same multi-view budget as the 2D models, video models remain the weakest in part-similarity. The vision-language pretrained Perception Encoder leads this group by a clear margin in `CKA`, providing limited evidence that multimodal supervision may also benefit part-level similarity in video encoders. V-JEPA and V-JEPA2 score comparatively well on `SVCCA` and `Mutual-kNN` despite their modest `CKA`, indicating greater disagreement between the three metrics within the video subset.

5.2 Do different language models rank vision models differently?

Our preceding analysis uses a fixed language encoder, raising the question of whether the observed vision-model rankings are specific to its representation geometry. We therefore rank a subset of 9 vision models with the previously used all-mpnet alongside 3 other language models: NV-Embed-v2 (Lee et al., 2024), CLIP Text (Radford et al., 2021), and Llama-3.1 (Grattafiori et al., 2024). The ranking induced under all models is highly correlated. Agreement is particularly high under `CKA` ($\rho \geq 0.88$), while `SVCCA` shows slightly greater sensitivity to the language model ($\rho \geq 0.78$). Nevertheless, the overall ranking remains stable, suggesting that our main findings are not specific to a single language encoder.

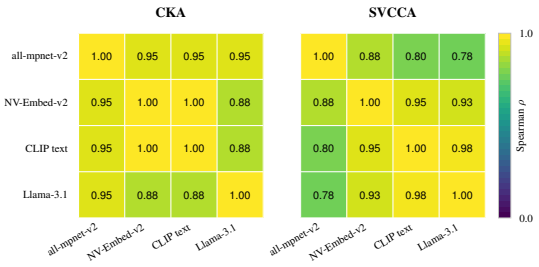


Figure 3: Language-model ranking agreement for 9 vision encoders under our part-similarity protocol. Each cell is the Spearman correlation (ρ) between the two language models' orderings of the vision encoders under CKA and SVCCA.

Findings
1. Part-similarity with language can emerge without multimodal pretraining.
2. Segmentation objectives may not produce stronger part-similarity with language.
3. Pretraining scale appears more consequential for part-similarity than the objective alone.

6 Part-Alignment

Practical vision-language systems rarely consume frozen features as-is; they typically insert a learned bridge such as a projector or adapter to bridge modalities. Thus, beyond similarity, what remains to be answered is how readily one modality can be mapped onto the other. As before we focus on 3D parts, a data-scarce regime where paired multimodal labels remain limited and costly to obtain. This motivates a probe that measures the *ease* of alignment under scarce supervision. We term this property part-alignment, and ask: *If part-level correspondences are unknown, which vision encoders can still be brought into correspondence with language models?* Below we show how such a probe is built and used to rank vision models.

6.1 Probing for Part-Alignment

We fix our probe to a 4-layer MLP f_a placed on top of frozen vision embeddings. Given visual features $F_i = f_v(x_i)$ from Section 3, we first apply the alignment head and then pool over each part mask

$$F'_i = f_a(F_i), \qquad \mathbf{z}'_{i,k} = \mathrm{Pool}(F'_i[Y_{i,k}]), \tag{1}$$

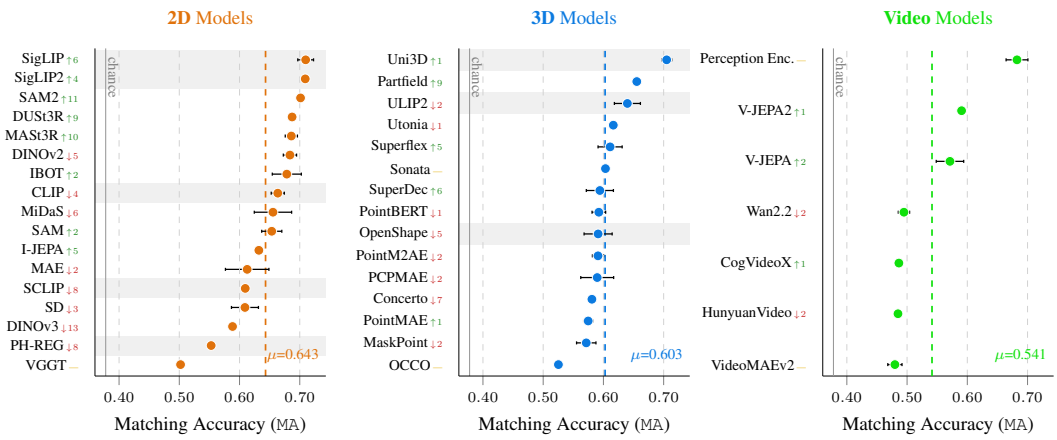


Figure 4: Part-alignment results grouped by modality. Each circle reports the mean across runs, with whiskers showing one standard deviation. Dashed lines mark modality means, while the solid gray line denotes chance performance. ↑ / ↓ / — denote rank change vs. part-similarity (Figure 2).

from which we form the pairwise cost matrix between projected visual parts and language embeddings,

$$\mathbf{C}_{i,pq} = c\big(\mathbf{z}'_{i,p}, \mathbf{e}_{i,q}\big)\,, \tag{2}$$

As in Equation (4), f_a is trained for one epoch to minimize the debiased Sinkhorn Divergence. We use the implementation from (Feydy et al., 2019), with $p = 2$ and $\texttt{blur} = 0.05$. The loss is computed separately between the visual-part and language-embedding sets of each instance and then averaged across the batch. Critically, this is what allows for training f_a without part-level correspondences.

At evaluation, part-alignment reduces to a matching problem for which we use the Hungarian algorithm (Kuhn, 1955) to find the optimal assignment between projected part features and language embeddings per instance obtained from the PartNet testing set.

$$\hat{\pi}_i = \arg\min_{\pi \in \mathfrak{S}_K} \sum_{p=1}^{K} \mathbf{C}_{i,p\pi(p)}, \quad \texttt{MA}_i = \frac{1}{K} \sum_{p=1}^{K} \mathbf{1}[\hat{\pi}_i(p) = p]\,. \tag{3}$$

where $\hat{\pi}_i$ denotes the recovered optimal permutation between vision and language embeddings, and $\texttt{MA}_i$ measure the fraction of correctly matched parts for instance i. In our experiments we report the final matching accuracy (MA) by first averaging $\texttt{MA}_i$ within each PartNet category and then average across all categories to account for class-imbalance. Note that we deliberately avoid re-measuring part-similarity (e.g. CKA) on the projected features $f_a(F_i)$ as the accuracy of labeling parts is easier to interpret with clearer bounds (chance level = $\frac{1}{K}$) while being better compatible with our training objective.

6.2 Does similarity predict alignment?

A natural question is whether the part-similarity rankings from Section 5 already foreshadow which encoders are easiest to map to language under our unpaired probe. We therefore compute Spearman's rank correlation between the within-modality orderings induced by CKA (Figure 2) and by matching accuracy (Figure 4). Across all models the correlation is only moderate (ρ=0.42), and it is driven almost entirely by 3D (ρ=0.58) and video (ρ=0.75) encoders; for 2D models the two rankings are essentially uncorrelated (ρ=0.05). In other words, high part-similarity does not reliably predict high part-alignment. This gap motivates treating alignment as a complementary criterion to similarity.

6.3 PART-ALIGNMENT ACROSS VISION MODELS AND MODALITIES

As shown in Figure 4, all evaluated vision encoders exceed chance-level matching accuracy, with the best encoders matching more that 70% of the parts accurately without being trained with visual part and language embedding correspondences. Notably, this includes the vision-only SAM 2 encoder. These results suggest that set-level supervision may support fine-grained semantic tasks in data-scarce settings such as 3D without requiring paired part annotations.

Rank changes relative to part-similarity reveal a consistent preference for spatially structured representations: PartField, SuperFlex, and SuperDec rise among 3D encoders, while SAM 2, MASt3R, and DUSt3R rise among 2D encoders. Conversely, OpenShape, Concerto, DINOv3, SCLIP, and PH-REG decline despite moderate or strong initial similarity. These results reinforce that part-similarity alone does not determine how readily an encoder can be mapped to language; visual representations that retain spatial or part-level structure are more amenable to alignment.

Findings
1. Strong part-similarity does not imply high part-alignment.
2. *Part-name correspondences can be learned from set-level supervision without known part-level pairs.*
3. *Spatially and part-aware visual representations are more readily aligned, irrespective of their initial part-similarity.*

7 PART-TRANSFERABILITY

As discussed in Section 2, 3D part segmentation and naming typically rely on paired part segments and labels, whether from human annotation or from expensive cross-modal labeling pipelines. Each adds annotation or compute cost and further limits how such fine-grained methods scale. Given that our probe can be trained without pairing, we ask: *can a mapping learned from unpaired, cheaply obtained part annotations transfer to unseen part hierarchies?*

7.1 EXPERIMENTAL SETUP

Training bags. To obtain unpaired training annotations, we derive part segments from SuperFlex (Tavernini et al., 2026), an unsupervised point-cloud decomposition method that segments objects at a geometric granularity and ranks highly on our part-alignment probe in Section 6. For part names, we keep the pipeline deliberately simple: we render each object from six fixed viewpoints and prompt InternVL2.5-8B (Chen et al., 2024) once to propose a bag of part names. We call this setting Unpaired in Table 1. We compare against a chance baseline and three oracles trained with different sources for segments and labels:

(i) Language Oracle : SuperFlex segments paired with ground-truth part names;
(ii) Segmentation Oracle : ground-truth segments paired with VLM part names;
(iii) Combined Oracle : ground-truth segments and part names.

Unpaired+ : Improving transferability. We further consider a variant of Unpaired with three simple changes that raise the ceiling for unpaired transfer:

1. We assume the part-label vocabulary is known *a priori* and expose this vocabulary to the VLM prompt.
2. Although our optimal-transport objective handles uneven bags, matching set cardinalities yields faster convergence. Specifically, we merge SuperFlex segments until their count equals the number of VLM-proposed part names, and adapt our loss to be more forgiving to merged segments by using a larger blur (0.2).
3. Unlike Section 6, where f_a is applied to dense features before pooling, we observe better generalization for SuperFlex segments by placing the alignment head on point features assigned to each discovered segment.

Evaluation. At test time we use the same matching accuracy (MA) as in the part-alignment study (Equation (3)): after training, we measure how accurately ground-truth segments are matched to part names from the test set. This isolates whether a probe trained only on unpaired discoveries transfers to human-defined part hierarchies, rather than fitting the SuperFlex-VLM training distribution alone. We additionally report two robustness metrics:

- MA_{hier}: evaluated across all part hierarchies obtained by successively merging the part tree in the PartNet test set for each instance;
- $\text{MA}_{\text{unseen}}$: evaluated on part hierarchies of categories held out during probe training.

7.2 Results

As shown in Table 1, training on SuperFlex segments and VLM labels (Unpaired) performs on par with the Segmentation Oracle and Language Oracle , and consistently exceeds chance across all metrics. With simple improvements (Unpaired+), we reduce the gap to the Combined Oracle , indicating that independently discovered geometries and labels can recover a large fraction of the transferable signal without paired annotations. The MA_{hier} results stay consistent with the standard MA, while in $\text{MA}_{\text{unseen}}$ unpaired transfer observes a significantly closer performance to the Combined Oracle .

Setting	Segments	Labels	MA	MA_{hier}	$\text{MA}_{\text{unseen}}$
Chance	–	–	37.8	40.7	38.1
Unpaired	SuperFlex	VLM	45.0	47.9	42.2
Unpaired+	SuperFlex	VLM	**50.7**	**54.5**	**47.7**
Oracles					
Language	SuperFlex	GT	42.2	44.3	39.6
Segmentation	GT	VLM	44.1	45.1	45.7
Combined	GT	GT	**61.1**	**63.2**	**52.9**

Table 1: Part-transferability on PartNet: MA, MA_{hier} & $\text{MA}_{\text{unseen}}$ after training the unpaired probe on independently discovered part segments and labels. Unpaired+ denotes the variant with access to the downstream part-label vocabulary during VLM prompting. GT refers to ground-truth segments/labels from PartNet.

Findings
1. *An unpaired probe trained on discovered segments and VLM part names transfers to human-defined PartNet hierarchies.*
2. *Simple training and prompting choices (Unpaired+) can increase transferability.*

8 Conclusion

In this paper, we investigate whether visual representations exhibit similarity with language at the part level. Towards this, we conduct a comprehensive analysis of part-similarity and part-alignment across 39 vision encoders spanning three modalities, and show that the two capture distinct properties: high part-similarity does not necessarily imply easy part-alignment. At the modality level, 2D encoders perform better on average than 3D and video encoders on both measures, although strong 3D models remain competitive. More broadly, our results show that pretraining strategy and architectural inductive biases strongly shape fine-grained vision-language organization, while explicit vision-language pretraining alone is not sufficient. Finally, we show that our unpaired probe transfers to unseen part hierarchies, enabling part matching between independently obtained bags of segments and part names without explicit paired annotations.

AI use statement

In this work, we used generative AI tools for drafting and editing text specifically for grammar correction, flagging nomenclature issues, and limited manuscript editing. Plotting programs are written with AI assistance. Majority of the implementation for experiments conducted are written and reviewed directly by the authors with only utilties writing and refactoring done with the help of AI tools. We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including text, claims, and artifacts.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture, April 2023. URL http://arxiv.org/abs/2301.08243. arXiv:2301.08243 [cs.CV].

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey, 2025. URL https://arxiv.org/abs/2404.18930.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. URL https://arxiv.org/abs/2404.08471.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

Yinjie Chen, Zipeng Yan, Chong Zhou, Bo Dai, and Andrew F. Luo. Vision Transformers with Self-Distilled Registers, November 2025. URL http://arxiv.org/abs/2505.21501. arXiv:2505.21501 [cs.CV].

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300, 2013.

Elisabetta Fedele, Boyang Sun, Leonidas Guibas, Marc Pollefeys, and Francis Engelmann. Superdec: 3d scene decomposition with superquadrics primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24625–24635, October 2025.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 2681–2690. PMLR, 2019.

Marco Garosi, Riccardo Tedoldi, Davide Boscaini, Massimiliano Mancini, Nicu Sebe, and Fabio Poiesi. 3D part segmentation via geometric aggregation of 2D visual features. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 3257–3267, February 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic Representation Hypothesis: An Aristotelian View, June 2026. URL http://arxiv.org/abs/2602.14486. arXiv:2602.14486 [cs.LG].

Souhail Hadgi, Luca Moschella, Andrea Santilli, Diego Gomez, Qixing Huang, Emanuele Rodolà, Simone Melzi, and Maks Ovsjanikov. Escaping Plato's Cave: Towards the Alignment of 3D and Text Latent Spaces, June 2025. URL https://arxiv.org/abs/2503.05283.

Souhail Hadgi, Bingchen Gong, Ramana Sundararaman, Emery Pierson, Lei Li, Peter Wonka, and Maks Ovsjanikov. Patchalign3d: Local feature alignment for dense 3d shape understanding, 2026. URL https://arxiv.org/abs/2601.02457.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, December 2021. URL http://arxiv.org/abs/2111.06377. arXiv:2111.06377 [cs.CV].

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic Representation Hypothesis, July 2024. URL https://arxiv.org/abs/2405.07987.

Jiachen Jiang, Jinxin Zhou, Bo Peng, Xia Ning, and Zhihui Zhu. Analyzing fine-grained alignment and enhancing vision understanding in multimodal language models. *Advances in Neural Information Processing Systems*, 38:874–899, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/017c897b4d85a744f345ccbf9d71e501-Abstract-Conference.html.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, Paris, France, October 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.00371. URL https://ieeexplore.ieee.org/document/10378323/.

A Sophia Koepke, Daniil Zverev, Shiry Ginosar, and Alexei A Efros. Back into Plato's Cave: Examining Cross-modal Representational Convergence at Scale. April 2026. doi: 10.48550/arXiv.2604.18572. URL https://arxiv.org/abs/2604.18572.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R, June 2024. URL http://arxiv.org/abs/2406.09756. arXiv:2406.09756 [cs.CV].

Haotian Liu, Mu Cai, and Yong Jae Lee. Masked Discrimination for Self-Supervised Learning on Point Clouds, August 2022. URL http://arxiv.org/abs/2203.11183. arXiv:2203.11183 [cs.CV].

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3D shape representation towards open-world understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44860–44879, 2023. doi: 10.52202/075280-1944. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c7304e77c832ddc70075dfee081ca6c-Abstract-Conference.html.

Minghua Liu, Mikaela Angelina Uy, Donglai Xiang, Hao Su, Sanja Fidler, Nicholas Sharp, and Jun Gao. PARTFIELD: Learning 3D Feature Fields for Part Segmentation and Beyond, April 2025. URL http://arxiv.org/abs/2504.11451. arXiv:2504.11451 [cs.CV].

Ziqi Ma, Yisong Yue, and Georgia Gkioxari. Find any part in 3d, 2024. URL https://arxiv.org/abs/2411.13550.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 909–918, June 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs.CV].

Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked Autoencoders for Point Cloud Self-supervised Learning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, volume 13662, pp. 604–621. Springer Nature Switzerland, Cham, 2022. ISBN 978-3-031-20085-4 978-3-031-20086-1. doi: 10.1007/978-3-031-20086-1_35. URL https://link.springer.com/10.1007/978-3-031-20086-1_35. Series Title: Lecture Notes in Computer Science.

Soumava Paul, Prakhar Kaushik, Ankit Vaidya, Anand Bhattad, and Alan Yuille. Name that part: 3d part segmentation and naming. In *CVPR Findings*, 2026. URL https://arxiv.org/abs/2512.18003. Accepted to CVPR Findings 2026.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. URL https://arxiv.org/abs/2306.14824.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, April 2017. URL http://arxiv.org/abs/1612.00593. arXiv:1612.00593 [cs.CV].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs.CV].

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188, October 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos, October 2024. URL http://arxiv.org/abs/2408.00714. arXiv:2408.00714 [cs.CV].

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It's a (blind) match! towards vision-language correspondence without parallel data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24983–24992, June 2025.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. Large vision-language model alignment and misalignment: A survey through the lens of explainability. *arXiv preprint arXiv:2501.01346*, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, August 2025. URL http://arxiv.org/abs/2508.10104. arXiv:2508.10104 [cs.CV].

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Gabriel Tavernini, Elisabetta Fedele, Tiago Novello, Leonidas Guibas, Marc Pollefeys, and Francis Engelmann. SuperFlex: Deformable Superquadrics for Point Cloud Decomposition, July 2026. URL http://arxiv.org/abs/2607.01015. arXiv:2607.01015 [cs.CV].

Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report, 2025. URL https://arxiv.org/abs/2511.18870.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, February 2025. URL http://arxiv.org/abs/2502.14786. arXiv:2502.14786 [cs.CV].

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking Self-Attention for Dense Vision-Language Inference, October 2024a. URL http://arxiv.org/abs/2312.01597. arXiv:2312.01597 [cs.CV].

Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matthew J. Kusner. Unsupervised Point Cloud Pre-Training via Occlusion Completion, October 2021. URL http://arxiv.org/abs/2010.01089. arXiv:2010.01089 [cs.CV].

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer, March 2025. URL http://arxiv.org/abs/2503.11651. arXiv:2503.11651 [cs.CV].

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560, June 2023.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy, December 2024b. URL http://arxiv.org/abs/2312.14132. arXiv:2312.14132 [cs.CV].

Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. In *CVPR*, 2024.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-Supervised Learning of Reliable Point Representations, March 2025. URL http://arxiv.org/abs/2503.16429. arXiv:2503.16429 [cs.CV].

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding, April 2024. URL http://arxiv.org/abs/2305.08275. arXiv:2305.08275 [cs].

Yunhan Yang, Yukun Huang, Yuan-Chen Guo, Liangjun Lu, Xiaoyang Wu, Edmund Y Lam, Yan-Pei Cao, and Xihui Liu. Sampart3d: Segment any part in 3d objects. *arXiv preprint arXiv:2411.07184*, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025. URL https://arxiv.org/abs/2408.06072.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-training 3D Point Cloud Transformers with Masked Point Modeling, June 2022. URL http://arxiv.org/abs/2111.14819. arXiv:2111.14819 [cs.CV].

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, Paris, France, October 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.01100. URL https://ieeexplore.ieee.org/document/10377550/.

Renrui Zhang, Ziyu Guo, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, Hongsheng Li, and Peng Gao. Point-M2AE: Multi-scale Masked Autoencoders for Hierarchical Point Cloud Pre-training, October 2022. URL http://arxiv.org/abs/2205.14401. arXiv:2205.14401 [cs.CV].

Xiangdong Zhang, Shaofeng Zhang, and Junchi Yan. PCP-MAE: Learning to Predict Centers for Point Masked Autoencoders, October 2024. URL http://arxiv.org/abs/2408.08753. arXiv:2408.08753 [cs.CV].

Yujia Zhang, Xiaoyang Wu, Yixing Lao, Chengyao Wang, Zhuotao Tian, Naiyan Wang, and Hengshuang Zhao. Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2338. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/649a31f2cb31a73b92c68b15bbf44442-Abstract-Conference.html.

Yujia Zhang, Xiaoyang Wu, Yunhan Yang, Xianzhe Fan, Han Li, Yuechen Zhang, Zehao Huang, Naiyan Wang, and Hengshuang Zhao. Utonia: Toward One Encoder for All Point Clouds, July 2026. URL http://arxiv.org/abs/2603.03283. arXiv:2603.03283 [cs.CV].

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer, January 2022. URL http://arxiv.org/abs/2111.07832. arXiv:2111.07832 [cs.CV].

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3D: Exploring Unified 3D Representation at Scale, October 2023. URL http://arxiv.org/abs/2310.06773. arXiv:2310.06773 [cs].

SUPPLEMENTARY MATERIAL

A PER-POINT FEATURE LIFTING

This section details how we obtain the modality-independent point features $F_i \in \mathbb{R}^{N_i \times d}$ introduced in Section 4.2. As in the main text, x_i denotes a PartNet mesh, $P_i = \{p_i^{(j)}\}_{j=1}^{N_i}$ its surface points, and f_v the visual encoder. 3D encoders operate directly on P_i. Image and video encoders instead produce dense feature maps that we back-project onto the same points, after which part pooling proceeds over $F_i[Y_{i,k}]$ as in Section 3.

3D Features. For 3D encoders, the point cloud P_i is processed directly by f_v. Since the encoder already operates on the object geometry, no feature lifting is required. The resulting per-point representation is simply

$$F_i = f_v(P_i),$$

where the j-th row of F_i is the feature associated with point $p_i^{(j)}$. For 3D encoders that operate on FPS-sampled point centroids (e.g. MaskPoint, Point-MAE, Point-BERT), we interpolate features back to all N_i points using inverse-distance-weighted kNN interpolation with $k = 3$.

Image Feature Lifting. For image encoders, f_v operates on rendered RGB views rather than directly on P_i. Let

$$\{I_{i,c}\}_{c=1}^{T}$$

denote a set of T rendered views of object x_i, indexed by virtual cameras $c = 1, \ldots, T$. Each view is processed independently to produce a dense feature map

$$M_{i,c} = f_v(I_{i,c}) \in \mathbb{R}^{H \times W \times d}.$$

For each rendered view, we precompute a pixel-to-point correspondence map

$$\pi_{i,c} : \Omega_{i,c} \to \{1, \ldots, N_i\} \cup \{\varnothing\},$$

where $\Omega_{i,c}$ denotes the pixel domain of view c (distinct from the 3D observation domain Ω_i in Section 3), $\pi_{i,c}(u) = j$ indicates that pixel u observes point $p_i^{(j)}$, and $\varnothing$ denotes background pixels.

The feature lifted from view c to point $p_i^{(j)}$ is

$$\mathbf{f}_{i,j,c} = \begin{cases} M_{i,c}(u), & \text{if } \pi_{i,c}(u) = j, \\ \mathbf{0}, & \text{otherwise.} \end{cases}$$

Since a point may be visible from multiple rendered views, the final point feature is obtained by averaging over all visible observations,

$$F_i[j] = \frac{1}{|\mathcal{C}_i(j)|} \sum_{c \in \mathcal{C}_i(j)} \mathbf{f}_{i,j,c},$$

where $F_i[j]$ denotes the j-th row of F_i and $\mathcal{C}_i(j)$ is the set of views in which $p_i^{(j)}$ is visible. Points that remain unobserved after multi-view aggregation are assigned features using k-nearest-neighbor interpolation on P_i. We verify in Section F that this multi-view averaging does not drive the part-similarity ranking of 2D encoders relative to a single-view baseline.

Video Feature Lifting. Video encoders follow the same lifting procedure, differing only in how the dense feature maps $M_{i,c}$ are computed. Instead of processing each view independently, f_v receives the ordered clip $\{I_{i,c}\}_{c=1}^{T}$ and jointly extracts temporally aware feature maps

$$M_{i,c} = f_v(I_{i,1:T}),$$

for each frame c. Although the encoder exploits temporal context, each frame still yields a dense spatial map over $\Omega_{i,c}$. These maps are back-projected with the same correspondences $\pi_{i,c}$ and aggregated with the same average over $\mathcal{C}_i(j)$, yielding the identical modality-independent representation $F_i \in \mathbb{R}^{N_i \times d}$ used for part pooling.

Table 2: Vision encoders evaluated in our part-similarity study (Figure 2), grouped by modality. VL denotes models with explicit vision–language pretraining.

Model	Modality	Reference
2D encoders		
VGGT	2D	(Wang et al., 2025)
I-JEPA	2D	(Assran et al., 2023)
MASt3R	2D	(Leroy et al., 2024)
SAM 2	2D	(Ravi et al., 2024)
DUSt3R	2D	(Wang et al., 2024b)
SAM	2D	(Kirillov et al., 2023)
SD	2D	(Rombach et al., 2022)
MAE	2D	(He et al., 2021)
iBOT	2D	(Zhou et al., 2022)
PH-REG	2D (VL)	(Chen et al., 2025)
SigLIP	2D (VL)	(Zhai et al., 2023)
SigLIP 2	2D (VL)	(Tschannen et al., 2025)
SCLIP	2D (VL)	(Wang et al., 2024a)
CLIP	2D (VL)	(Radford et al., 2021)
MiDaS	2D	(Ranftl et al., 2021)
DINOv3	2D	(Siméoni et al., 2025)
DINOv2	2D	(Oquab et al., 2024)
3D encoders		
OCCO	3D	(Wang et al., 2021)
Point-MAE	3D	(Pang et al., 2022)
SuperDec	3D	(Fedele et al., 2025)
MaskPoint	3D	(Liu et al., 2022)
PartField	3D	(Liu et al., 2025)
SuperFlex	3D	(Tavernini et al., 2026)
PCP-MAE	3D	(Zhang et al., 2024)
Point-M2AE	3D	(Zhang et al., 2022)
Point-BERT	3D	(Yu et al., 2022)
Sonata	3D	(Wu et al., 2025)
Concerto	3D	(Zhang et al., 2025)
OpenShape	3D (VL)	(Liu et al., 2023)
Utonia	3D	(Zhang et al., 2026)
Uni3D	3D (VL)	(Zhou et al., 2023)
ULIP-2	3D (VL)	(Xue et al., 2024)
Video encoders		
VideoMAE v2	Video	(Wang et al., 2023)
CogVideoX	Video	(Yang et al., 2025)
V-JEPA	Video	(Bardes et al., 2024)
HunyuanVideo	Video	(Team, 2025)
V-JEPA 2	Video	(Assran et al., 2025)
Wan 2.2	Video	(Wan et al., 2025)
Perception	Video	(Bolya et al., 2025)

B Vision Model Zoo

Table 2 lists every vision encoder used in our part-similarity experiments (Figure 2), grouped by modality and annotated when pretrained with explicit vision–language supervision. Citations are collected here so that the main text can discuss models by name directly.

C Centered Kernel Alignment

We compute `CKA` by constructing pairwise similarity matrices over paired visual and language part representations:

$$\mathbf{S}_{v,pq} = s(\mathbf{z}_p, \mathbf{z}_q)\,, \qquad \mathbf{S}_{l,pq} = s(\mathbf{e}_p, \mathbf{e}_q)\,,$$

where $s(\cdot, \cdot)$ typically denotes cosine similarity, p and q index a pair of parts, and $\mathbf{S}_v$ and $\mathbf{S}_l$ are the similarity matrices for the visual and language representations, respectively. The final CKA score is

$$\texttt{CKA}(\mathbf{S}_v, \mathbf{S}_l) = \frac{\mathrm{HSIC}(\mathbf{S}_v, \mathbf{S}_l)}{\sqrt{\mathrm{HSIC}(\mathbf{S}_v, \mathbf{S}_v)\,\mathrm{HSIC}(\mathbf{S}_l, \mathbf{S}_l)}},$$

where HSIC is the Hilbert–Schmidt Independence Criterion (Gretton et al., 2005),

$$\mathrm{HSIC}(\mathbf{S}_v, \mathbf{S}_l) = \frac{1}{(n-1)^2}\,\mathrm{Tr}(\mathbf{S}_v \mathbf{H} \mathbf{S}_l \mathbf{H}).$$

Here n is the sample size, $\mathrm{Tr}(\cdot)$ denotes the matrix trace, and $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is the centering matrix.

D Optimal Transport

This section formalizes the Optimal Transport (OT) and differentiable OT objectives used in the main text (Sections 3 and 6). Given visual part features $\{\mathbf{z}_p\}_{p=1}^{K_i}$ and language embeddings $\{\mathbf{e}_q\}_{q=1}^{K_i}$, define the pairwise cost matrix

$$\mathbf{C}_{pq} = c(\mathbf{z}_p, \mathbf{e}_q)\,. \tag{4}$$

Classical OT seeks a transport plan that minimizes the total cost

$$\min_{\mathbf{P} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle. \tag{5}$$

Here P_{pq} is the probability mass moved from visual part p to language concept q, and the marginals $\mathbf{a}, \mathbf{b} \in \Delta^{K_i - 1}$ are probability vectors such that

$$\Pi(\mathbf{a}, \mathbf{b}) = \{\mathbf{P} \geq 0 : \mathbf{P}\mathbf{1} = \mathbf{a},\ \mathbf{P}^\top \mathbf{1} = \mathbf{b}\}$$

is the set of admissible transport plans.

Differentiable Optimal Transport. To optimize transport plans end-to-end, we use entropy-regularized OT

$$\min_{\mathbf{P} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P}), \tag{6}$$

where

$$H(\mathbf{P}) = -\sum_{p,q} P_{pq} \log P_{pq} \tag{7}$$

is the entropy of the transport plan. The resulting problem can be solved efficiently with the Debiased Sinkhorn divergence (Ramdas et al., 2017; Feydy et al., 2019), which alternates row and column normalizations until convergence. In Section 6 we train the alignment head f_a by minimizing this divergence between projected visual-part sets and language-embedding sets.

E Global vs. Local Similarity

In Section 5 we rank vision encoders by global part-similarity to language using CKA and SVCCA (Figure 2). Those metrics compare relational structure over all sampled part pairs; they do not ask whether local neighborhoods in vision and language agree. We therefore report Mutual-kNN ($k = 10$) under the same modality-agnostic protocol as a complementary, neighborhood-level view. Absolute scores remain small (group means between 0.024 and 0.058), as expected when $k = 10$ over several thousand parts, so the metric is most informative comparatively.

Figure 5 preserves the broad modality ordering of Figure 2 (2D strongest, video weakest), but reshuffles several within-group ranks. Among 3D encoders, ULIP-2 leads under CKA yet is only mid-pack under Mutual-kNN, while Uni3D leads local agreement and Sonata/Concerto drop sharply; among 2D encoders, DINOv2 remains on top, but I-JEPA rises from near the bottom under CKA to among the strongest locally. On video, Perception Encoder leads both metrics, whereas V-JEPA and V-JEPA 2 score comparatively well on Mutual-kNN. Global subspace similarity and local neighborhood preservation therefore capture related yet distinct aspects of part-level vision-language similarity.

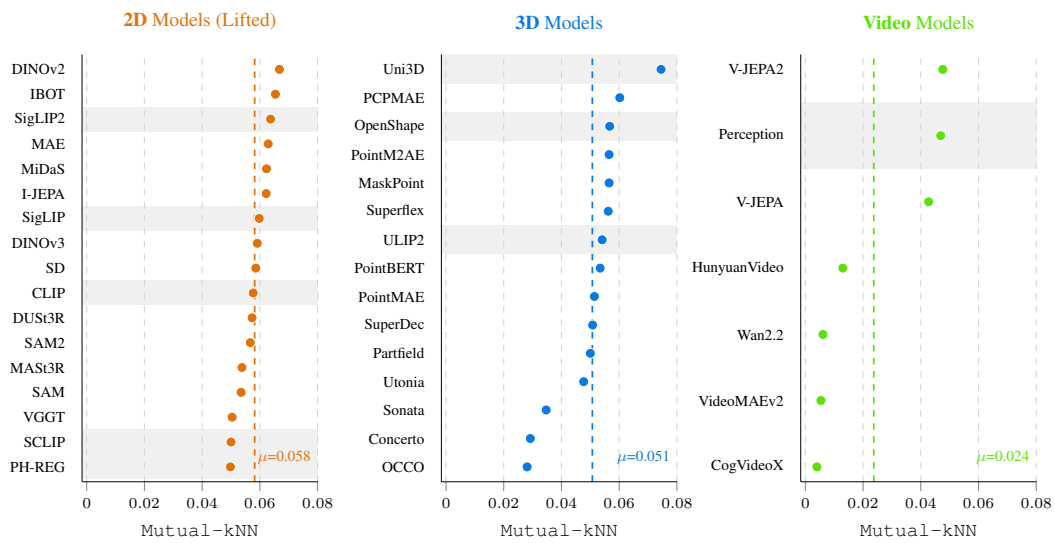


Figure 5: Part-Similarity using `Mutual-kNN` (k=10) for different visual models, grouped by modality. Shaded rows indicate models with explicit multi-modal pretraining. Companion to the global `CKA`/`SVCCA` ranking in Figure 2.

F Robustness to Lifting

Our default protocol lifts 2D features across multiple views and averages observations per point. To check whether this multi-view pooling drives the ranking of vision encoders, we recompute part–language `CKA` using features from a single rendered view of each PartNet shape and compare against the lifted multi-view scores. Figure 6 shows that single-view and lifted 2D `CKA` are strongly correlated across models (Pearson r=0.968, p=2.3×10^{-10}), so encoder rankings remain largely unchanged. Lifted scores are typically slightly higher than their single-view counterparts, consistent with denser coverage of the surface, but the relative ordering is preserved.

G Does Part-Alignment Probe Increase Similarity?

The main text evaluates the part-alignment probe with matching accuracy (`MA`; Section 6). Here we ask a complementary question: does training the probe also raise frozen-style part-similarity metrics on the projected features? We use PartField as a representative encoder with modest part-similarity under `CKA`/`SVCCA` (Figure 2) and measure `CKA`, `SVCCA`, and `Mutual-kNN` before and after training f_a. In the aligned setting, the probe is trained with ground-truth part masks and tags, with the projection applied before feature consolidation.

Encoder	Alignment	`CKA` ↑	`SVCCA` ↑	`Mutual-kNN` ↑
PartField	✗	0.165	0.244	0.050
PartField	✓	0.734	0.640	0.065

Table 3: Part-similarity of PartField features before and after training the part-alignment probe. Higher is better for all metrics.

As shown in Table 3, part-similarity rises substantially after probe training (`CKA`: 0.165→0.734). While these gains confirm that the learned mapping brings visual parts closer to language, we still prefer `MA` in the main text as it is more interpretable, has a clear chance baseline ($1/K$), and aligns more directly with the matching objective.

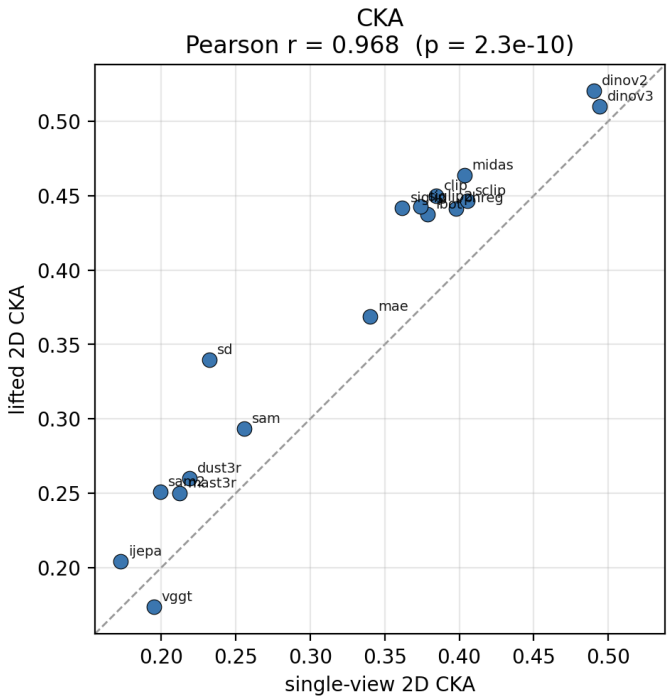


Figure 6: Single-view vs. lifted multi-view part-similarity CKA for 2D vision encoders on PartNet. Encoder rankings are highly consistent between the two protocols (Pearson r=0.968).

H Projection Location

The main-text probe places a 4-layer MLP f_a on frozen vision features and leaves language embeddings fixed (Section 6). Although this is our default, we also ask whether the reverse mapping (language $\rightarrow$ vision), or a shared projected space with heads on both sides, is equally easy under the same learning budget. We again use PartField as a representative encoder and report matching accuracy for three locations of the probe.

Encoder	**Projector Location**	MA ↑
PartField	Vision	**65.5**
PartField	Language	55.2
PartField	Both	38.0

Table 4: Effect of projection location on PartField part-alignment (MA).

Table 4 shows that mapping vision into language remains easiest. Transforming text into vision space under the same budget is noticeably harder, and jointly projecting both modalities into a shared space is harder still. We interpret this as evidence that language provides a more stable target for the OT objective: the diverse visual part embeddings likely provide a more diverse learning signal, rather than when the same part-label embedding must also be reshaped to varied geometrical embeddings or when both sides are free to transform.