

# Name That Part: 3D Part Segmentation and Naming

Soumava Paul<sup>★</sup> Prakhar Kaushik<sup>★‡</sup>

Ankit Vaidya Anand Bhattad Alan Yuille

★ Equal Contribution ‡ Project Lead

Johns Hopkins University  
Baltimore, MD, USA

{spaul27,pkaushi1,avaidya7,bhattad,ayuille1}@jh.edu

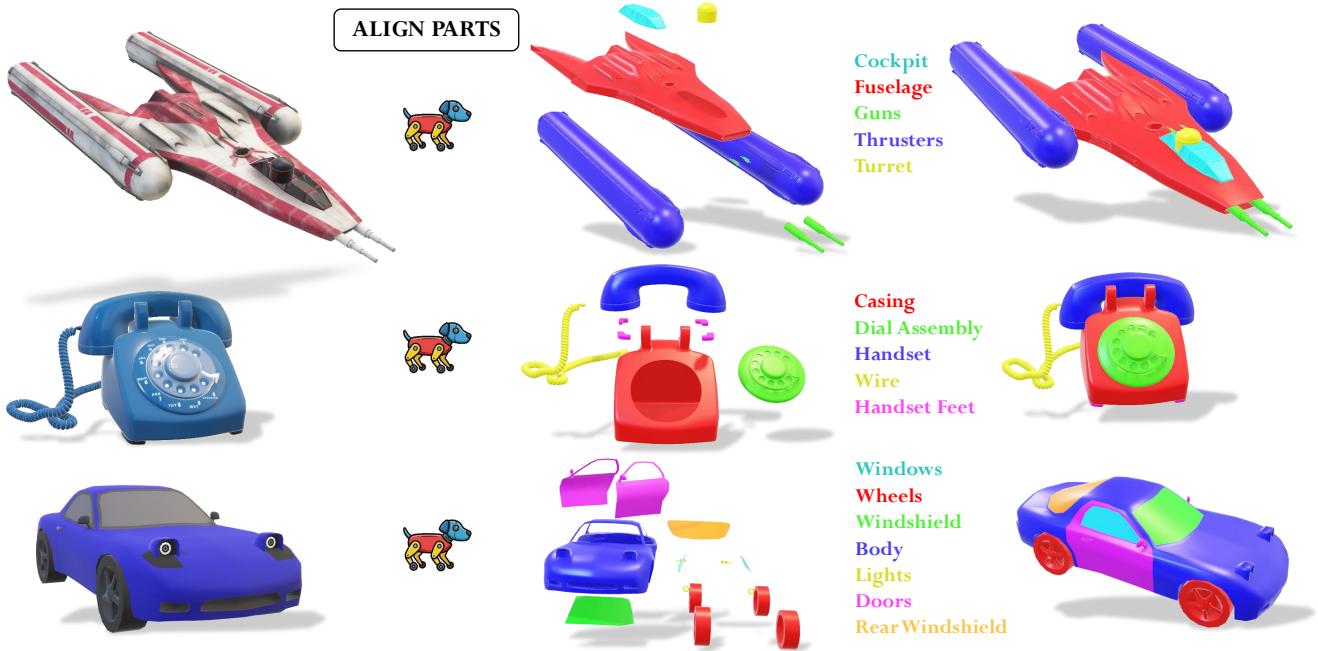


Figure 1: ALIGN-Parts is the first large-scale method to be able to efficiently and semantically segment and name 3D parts of an object, unlike previous methods, which could only do class-agnostic part segmentation. Our method is also feed-forward, and defines 3D parts according to human-oriented, object-specific affordance descriptions.

## Abstract

We address semantic 3D part segmentation: decomposing objects into parts with meaningful names. While datasets exist with part annotations, their definitions are inconsistent across datasets, limiting robust training. Previous methods produce unlabeled decompositions or retrieve single parts without complete shape annotations. We propose ALIGN-Parts, which formulates part naming as a direct set alignment task. Our method decomposes shapes into partlets - implicit 3D part representations - matched to part descriptions via bipartite assignment. We combine geometric cues from 3D part fields, appearance from multi-view vision features, and semantic knowledge from language-model-generated affordance descriptions. Text-alignment loss ensures partlets share embedding space with text, enabling a theoretically open-vocabulary matching setup, given sufficient data. Our efficient and novel, one-shot, 3D

part segmentation and naming method finds applications in several downstream tasks, including serving as a scalable annotation engine. As our model supports zero-shot matching to arbitrary descriptions and confidence-calibrated predictions for known categories, with human verification, we create a unified ontology that aligns PartNet, 3DCoMPaT++, and Find3D, consisting of 1,794 unique 3D parts. We also show examples from our newly created Tex-Parts dataset. We also introduce 2 novel metrics appropriate for the named 3D part segmentation task.

## Keywords

3D part semantic segmentation, geometric deep learning

## 1 Introduction

Many vision tasks require 3D parts, not just whole-object labels. Examples include robots grasping handles and creators editing assets. This requires solving two problems simultaneously: geometrically

Method	Complete Decomposition	Named Parts	Open Vocabulary	Permutation Invariant	Part-level Features	Feed Forward
PartField	✓	✗	✗	-	✗	✗
SAMPart3D	✗	✗	✗	✗	✗	✗
Find3D	✗	✓	✓	✗	✗	✗
<b>Ours</b>	✓	✓	✓	✓	✓	✓

Table 1: Comparison with 3D part methods.

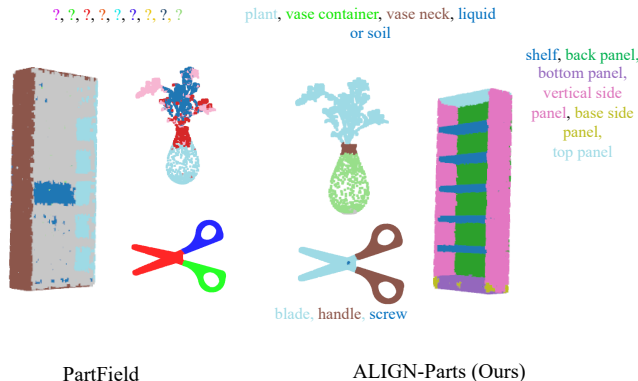


Figure 2: ALIGN-Parts segments and names 3D parts, unlike PartField [Liu et al. 2025], which only segments. Our method is 100× (post data pre-processing) faster at generating these segments along with names compared to PartField.

segmenting the parts and semantically naming them. While large datasets of 3D objects exist, only a few provide part annotations, and these annotations are often inconsistent across datasets [Ma et al. 2024; Mo et al. 2019; Slim et al. 2023]. An algorithm that can provide accurate and consistent annotations of named parts on any 3D object would enable scalable training data and support human-in-the-loop annotation pipelines.

Existing methods address only one aspect of this problem. Part segmentation models can identify geometric boundaries but produce unnamed regions [Liu et al. 2025]. Language-grounded systems can retrieve a single part from a text query but fail to produce a complete set of names for all parts of an object [Ma et al. 2024]. Classical unsupervised part discovery lacks the semantic grounding to derive consistent definitions. This gap has created a bottleneck: the absence of large-scale, consistently-annotated 3D part data. Inducing consistent parts from unlabeled web assets requires coupling geometric features and semantic knowledge with human verification.

We propose ALIGN-Parts, which formulates 3D part naming as a direct set alignment problem. Rather than deciding per-point which text label to assign, we decompose the shape into a small set of shape-conditioned *partlets*. Each partlet consists of a set of points (a segmentation mask) and a text description (embedding) corresponding to one part. These partlets aggregate information across all points belonging to a part: a single point on a chair seat contains limited information, but the set of all points on the seat

specifies the part. We align this set of partlets to a set of candidate part descriptions via bipartite matching in a single forward pass. Each partlet inherits a name from its matched description. A “null” class allows the model to discard unused partlets, enabling the number of parts to adapt per shape while ensuring permutation consistency: each predicted part receives at most one name, and each name is used at most once.

Our key contribution is applying set-level matching to 3D part fields. This enables lower computational complexity: we match a handful of partlets to descriptions instead of all points to all labels. It provides permutation consistency: each predicted part receives at most one name, and each name is used at most once. Because of this, ALIGN-Parts is 100× faster at generating these segments along with names.

To ensure partlets are both geometrically separable and semantically meaningful, we combine three signals. (1) Geometry features from a 3D part-field backbone [Liu et al. 2025] capture shape structure. (2) Appearance features from multi-view image encoders [Oquab et al. 2023] lifted onto 3D geometry provide texture cues.

(3) Semantic knowledge comes from affordance-aware part descriptions that encode form-and-function relationships [Connell and Brady 1987]. For example, a chair seat becomes “the horizontal surface where a person sits,” linking its flat, horizontal geometry to its sitting affordance.

For text grounding, we generate these descriptions using a large language model [Comanici et al. 2025] and embed them using MP-Net sentence transformers [Song et al. 2020]. By representing part descriptions as embeddings in a continuous space, our approach supports scenarios where the model can match partlets to any user-provided set of part descriptions without retraining.

Our approach provides a tool to address the data bottleneck. We construct a unified part ontology using a hybrid LLM-and-human process that normalizes labels and verifies geometric consistency across PartNet, 3DCoMPaT++, and Find3D. We then apply ALIGN-Parts to bootstrap annotations from unlabeled TexVerse assets, creating Tex-Parts: a dataset with 8450 objects spanning 14k part categories. ALIGN-Parts serves as a *scalable annotation engine* that proposes named parts for human verification, converting raw meshes into training data with minimal effort and enabling the construction of large, consistently-annotated 3D datasets.

In summary, our contributions are:

- **Direct 3D parts alignment for open-world part naming.** We introduce *partlets* - shape-conditioned part proposals with text embeddings - and match them to candidate descriptions via bipartite assignment, enabling efficient labeling of 3D part segmentation and naming.

- **Geometry-appearance-semantic fusion.** We combine geometric structure, appearance features, and affordance-aware LLM-generated descriptions to produce semantically grounded, visually coherent partlets.
- **Metrics for evaluating 3D part semantic segmentation** We introduce 2 metrics suitable for our 3D part segmentation and naming task, which evaluate ALIGN-Parts and related baselines for both part segmentation accuracy and semantic correctness of the predicted part label.
- **Unified ontology and scalable annotation engine.** We harmonize part taxonomies across PartNet, 3DCoMPaT++, and Find3D, and demonstrate a human-in-the-loop pipeline that bootstraps **TexParts**, a verified benchmark of 8450 objects spanning 14k categories derived from Texverse [Zhang et al. 2025]. ALIGN-Parts converts raw meshes into training data with 5-8× less human effort than manual annotation.

## 2 Related Work

**3D Part Segmentation.** Traditional methods operate in a purely geometric regime. PartField [Liu et al. 2025], the current state-of-the-art, learns dense per-point feature fields but produces unlabeled regions. Recent works lift 2D foundation models into 3D: SAM-based approaches [Ma et al. 2025; Tang et al. 2024; Yang et al. 2024] adapt Segment-Anything via multi-view projection but require prompts and lack semantic names. PartSTAD [Kim and Sung 2024] adapts GLIP and SAM with 3D-aware objectives, while Diff3F [Dutt et al. 2024] exploits diffusion features, but both operate as multi-stage pipelines without producing complete, non-overlapping named parts.

To our knowledge, [Kalogerakis et al. 2010, 2017] are the only previous works that do simultaneous 3D part segmentation and labeling. [Kalogerakis et al. 2010] defines the task by assigning fixed labels to mesh faces using statistical consistency using a Conditional Random Field Model, which is hard to scale up. Further, the model lacks any semantic knowledge given its reliance on hand engineering features and geometric consistency. Simultaneous 3D part segmentation and recognition has remained a complex task with little progress since then as it requires incorporation of human contextual knowledge of parts. Therefore, most subsequent works have focused on amodal segmentation, which is a relatively easier task.

Other important works in the area of 3D parts - which are not directly relevant to our work - but involve relevant 3D analysis, understanding or retrieval, include [Golovinskiy and Funkhouser 2008; Hanocka et al. 2019; Jones et al. 2020; Shapira et al. 2008; Stelzner et al. 2021; Takmaz et al. 2023; Tulsiani et al. 2017].

**Language-Grounded 3D Understanding.** PartSLIP [Liu et al. 2023], PartSLIP++ [Zhou et al. 2023], and PartDistill [Umam et al. 2024] use image-language models for part segmentation but require per-category fine-tuning with predefined vocabularies. Find3D [Ma et al. 2024] is most related: it trains a point transformer in text embedding space for text-to-part retrieval. However, Find3D operates query-by-query (given "wing, head," it returns masks) rather than producing complete decompositions, and works on individual point features rather than part-level aggregations.

**ALIGN-Parts differs fundamentally.** We formulate part segmentation and naming as *direct set alignment*: Partlets aggregate point features into part-level representations matched to text via optimal transport, enabling (1) complete, non-overlapping decompositions in one pass, (2) permutation-consistent naming, (3) dynamic part counts, and (4) zero-shot generalization (Table 1). Unlike previous methods, ALIGN-Parts jointly learns segmentation and alignment in a single feed-forward pass. Part count emerges automatically from activated Partlets, eliminating dependence on K prediction or part-name inputs.

**Part-Based Generation and Datasets.** Generative methods [Chen et al. 2024, 2025; Lin et al. 2025; Tang et al. 2025; Yang et al. 2025] discover part structure through synthesis but lack open-vocabulary naming mechanisms. For datasets, PartNet [Mo et al. 2019], 3DCoMPaT++ [Slim et al. 2023], and GPartNet [Geng et al. 2023] provide part annotations but use inconsistent taxonomies. ALIGN-Parts constructs a unified ontology across these datasets and creates a unified dataset for cross-dataset evaluation.

## 3 Method

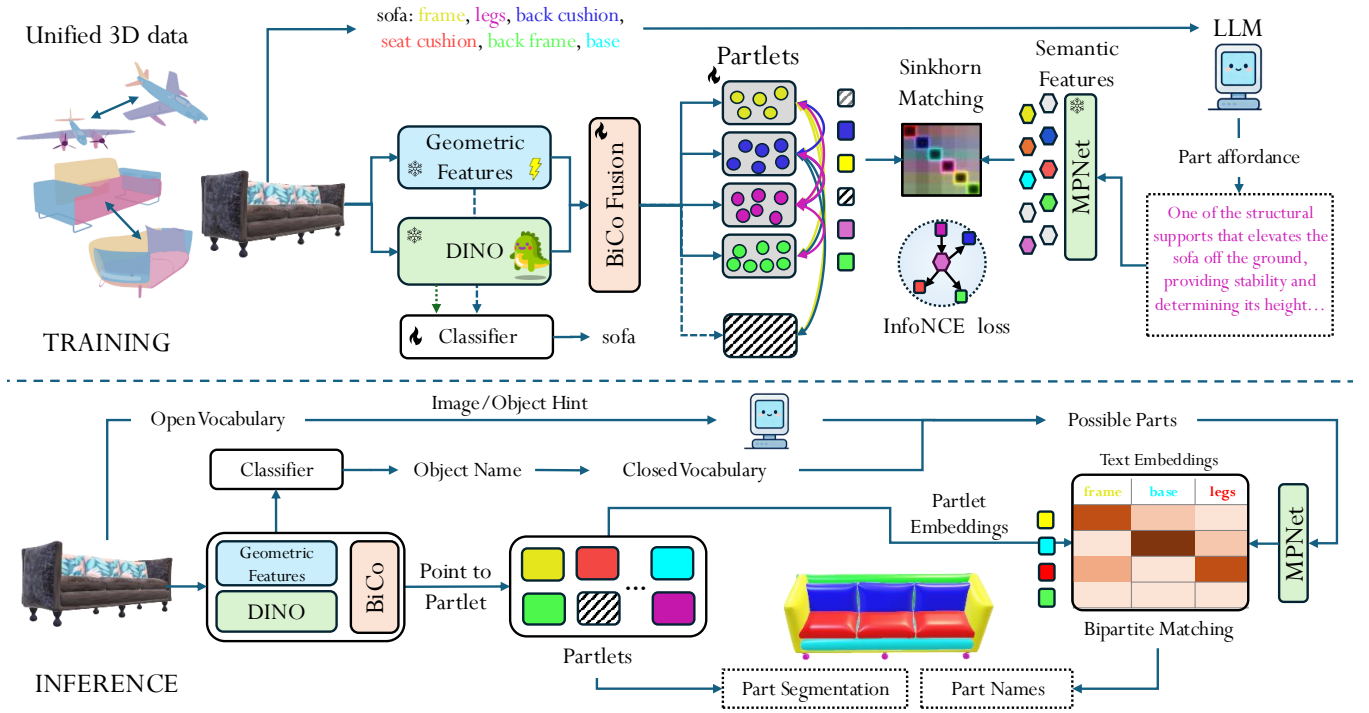
**Overview.** We propose ALIGN-Parts, a framework that treats automatic semantic 3D part segmentation as a direct set alignment problem, analogous to DETR [Carion et al. 2020] in 2D detection. The pipeline consists of three components: (1) a dense feature fusion module combining geometry and appearance via localized bi-directional 3D-aware cross-attention, (2) a Partlets module that learns K adaptive part-level representations by aggregating fused point features, and (3) a semantic grounding module that aligns Partlets to part names via text descriptions of part affordances. At inference, given a 3D shape and candidate part descriptions, our method produces named part segments without specifying part counts, names, or point prompts. Training uses contrastive alignment and differentiable optimal transport matching between predicted Partlets and ground-truth parts, enabling end-to-end learning of both segmentation and semantics.

**Problem Formulation.** Given a 3D point cloud  $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^N$ , each point has geometry-first features  $\mathbf{f}_i^g \in \mathbb{R}^{d_g}$ , appearance-first features  $\mathbf{f}_i^a \in \mathbb{R}^{d_a}$ , and coordinates  $\mathbf{x}_i \in \mathbb{R}^3$ . During training, we have ground-truth part masks  $\mathbf{M}^{\text{gt}} \in \{0, 1\}^{A \times N}$  and corresponding text embeddings  $\{\mathbf{t}_a\}_{a=1}^A$  where  $\mathbf{t}_a \in \mathbb{R}^{d_t}$  is an MPNet embedding of a language-model-generated part description.

Our model learns  $K = 32$  instance-specific Partlet embeddings  $\{\mathbf{s}_k \in \mathbb{R}^{d_t}\}_{k=1}^K$  that reside in the same semantic space as text embeddings (we use  $d_t = 768$ ). Each Partlet predicts three outputs: (i) a soft mask  $\mathbf{m}_k \in [0, 1]^N$  over points, (ii) a partness score  $\text{part}_k \in \mathbb{R}$  indicating whether it represents an actual part, and (iii) its embedding  $\mathbf{s}_k$  which serves as the prototype for semantic alignment. We select  $K = 32$  to accommodate variable part configurations (most shapes have  $\leq 28$  parts) while allowing future extensibility. Point labels are obtained by matching Partlets to ground-truth parts via optimal transport.

### 3.1 Architecture

**Dense Feature Fusion Module** The raw geometric features  $\mathbf{f}_i^g$  and appearance features  $\mathbf{f}_i^a$  capture complementary information: geometry encodes shape characteristics, while appearance provides



**Figure 3: ALIGN-Parts.** Overview of the ALIGN-Parts framework for language-grounded 3D part segmentation and naming. *Top: training.* Given a 3D from our semantically unified 3D parts data, geometry features are extracted with PartField and appearance features with DINOv2 from multi-view renderings; these are fused by the BiCo Fusion module using efficient bi-directional cross-attention on local  $k=16$  nearest-neighbor graphs in 3D space, reducing complexity from  $O(N^2)$  to  $O(Nk)$  and yielding enriched point features. A decoder then learns  $K$  part-level “Partlet” representations that aggregate the fused features, with segmentation supervision provided at the Partlet level. To semantically ground Partlets, an LLM generates affordance-aware descriptions for each possible part (e.g., “structural supports that elevate the sofa” for sofa legs), which are embedded by a pretrained MPNet encoder; Sinkhorn matching establishes a bipartite assignment between Partlet and text embeddings, and an InfoNCE loss further aligns them in a shared representation space while a classifier predicts the object category. *Bottom: inference.* At test time, ALIGN-Parts operates in both closed-vocabulary (object categories similar to those seen in training) and open-vocabulary (novel object categories) settings: in the closed-vocabulary case, the trained 3D classifier predicts the object class and retrieves its candidate part list, whereas in the open-vocabulary case an LLM proposes an overcomplete set of plausible parts for the queried object. Given these candidate part texts, their MPNet embeddings are bipartite-matched to the predicted Partlets, which jointly produce 3D part segmentation masks and corresponding part names.

texture and visual cues. We fuse these modalities through a bi-directional cross-attention module that operates on a  $k$ -nearest neighbor (KNN) graph to maintain computational tractability. Given the KNN graph structure with indices  $\mathcal{N}_i$  denoting the  $k$ -nearest neighbors of point  $i$ , we compute:

*Relative Positional Bias.* To incorporate 3D spatial structure through Fourier-encoded relative positional biases, for each neighbor pair  $(i, j)$  where  $j \in \mathcal{N}_i$ , we compute:

$$\mathbf{d}_{ij} = \mathbf{x}_j - \mathbf{x}_i \in \mathbb{R}^3 \quad (1)$$

$$\mathbf{f}_{ij} = \mathbf{d}_{ij} \odot \omega \in \mathbb{R}^{3 \times F} \quad (2)$$

$$\mathbf{h}_{ij} = [\sin(\mathbf{f}_{ij}), \cos(\mathbf{f}_{ij})] \in \mathbb{R}^{3 \times 2F} \quad (3)$$

$$\mathbf{b}_{ij} = \text{MLP}([\mathbf{d}_{ij}, \text{flatten}(\mathbf{h}_{ij})]) \in \mathbb{R}^H \quad (4)$$

where  $\omega = [2^0, 2^1, \dots, 2^{F-1}]$  with  $F=6$  are logarithmically-spaced frequencies, and the MLP consists of two layers:  $\mathbb{R}^{39} \rightarrow \mathbb{R}^{64} \rightarrow \mathbb{R}^H$  with ReLU activation.

*Bi-Directional Cross-Attention.* Let  $H = 8$  be the number of attention heads and  $d_h = d_m/H = 96$  be the head dimension where  $d_m = 768$  is the model dimension.

**Geometric-to-Appearance Direction:** For each point  $i$  with neighbors  $N_i$ , geometric features attend to appearance features of neighbors, producing cross-modal features that capture appearance information.

$$\mathbf{Q}_p^i = \mathbf{W}_q^p \mathbf{f}_i^g \in \mathbb{R}^{H \times d_h} \quad (5)$$

$$\mathbf{K}_a^{ij} = \mathbf{W}_k^a \mathbf{f}_j^a \in \mathbb{R}^{H \times d_h}, \quad \forall j \in N_i \quad (6)$$

$$\mathbf{V}_a^{ij} = \mathbf{W}_v^a \mathbf{f}_j^a \in \mathbb{R}^{H \times d_h}, \quad \forall j \in N_i \quad (7)$$

$$\alpha_{pa}^{i,h,j} = \frac{\exp((\mathbf{Q}_p^i[h] \cdot \mathbf{K}_a^{ij}[h]) / \sqrt{d_h} + \mathbf{b}_{ij}[h])}{\sum_{j' \in N_i} \exp((\mathbf{Q}_p^i[h] \cdot \mathbf{K}_a^{ij'}[h]) / \sqrt{d_h} + \mathbf{b}_{ij'}[h])} \quad (8)$$

$$\mathbf{z}_p^{i,h} = \sum_{j \in N_i} \alpha_{pa}^{i,h,j} \mathbf{V}_a^{ij}[h] \quad (9)$$

$$\mathbf{r}_p^i = \mathbf{W}_{pa} \text{concat}_h[\mathbf{z}_p^{i,h}] \in \mathbb{R}^{d_g} \quad (10)$$

**Appearance-to-Geometric Direction:** Symmetrically, appearance features attend to geometric features of neighbors, capturing geometric information. This produces  $\mathbf{r}_a^i \in \mathbb{R}^{d_a}$ .

**Gated Fusion.** Learned sigmoid gates control how much of this cross-modal information to incorporate into each original feature, based on both the original feature and the attended information.

$$\mathbf{g}_p^i = \sigma(\mathbf{W}_g^p [\mathbf{f}_i^g; \mathbf{r}_p^i]) \in \mathbb{R}^{d_g} \quad (11)$$

$$\mathbf{g}_a^i = \sigma(\mathbf{W}_g^a [\mathbf{f}_i^a; \mathbf{r}_a^i]) \in \mathbb{R}^{d_a} \quad (12)$$

$$\tilde{\mathbf{f}}_i^g = \text{LayerNorm}(\mathbf{f}_i^g + \mathbf{g}_p^i \odot \mathbf{r}_p^i) \quad (13)$$

$$\tilde{\mathbf{f}}_i^a = \text{LayerNorm}(\mathbf{f}_i^a + \mathbf{g}_a^i \odot \mathbf{r}_a^i) \quad (14)$$

**Final Projection.** After gated addition and layer normalization, we concatenate both modalities and project through a two-layer MLP to produce fused features  $\mathbf{h}_i \in \mathbb{R}^{1216}$  for each point.

$$\mathbf{h}_i = \mathbf{W}_2 \text{GELU}(\mathbf{W}_1 \text{LayerNorm}([\tilde{\mathbf{f}}_i^g; \tilde{\mathbf{f}}_i^a])) \quad (15)$$

where  $\mathbf{W}_1 \in \mathbb{R}^{d_f \times (d_g + d_a)}$  and  $\mathbf{W}_2 \in \mathbb{R}^{d_f \times d_f}$  with  $d_g = 448$ ,  $d_a = 768$ , and  $d_f = 256$ . BiCo Fusion operates on local  $k=16$  nearest neighbor graphs in 3D coordinate space, reducing complexity to  $\mathcal{O}(Nk)$ .

**Points  $\rightarrow$  Partlets: Learning Part-Level Representations.** We learn  $K$  Partlet embeddings that aggregate point-level information into part-level representations. A Partlet is defined by three components: (i) a soft segmentation mask  $\mathbf{m} \in [0, 1]^N$  representing membership scores, (ii) a Partlet embedding  $\mathbf{s} \in \mathbb{R}^{d_t}$  in the learned semantic space, and (iii) a text embedding  $\mathbf{z} \in \mathbb{R}^{d_t}$  representing the part description.

Formally, we learn a parameterized function  $f_\theta : \mathbb{R}^{d_h \times N} \rightarrow (\mathbb{R}^N \times \mathbb{R}^{d_t})^K$  that maps fused point features  $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^N$  to  $K$  Partlets:

$$(\mathbf{m}_k, \mathbf{s}_k) = f_\theta(\mathbf{H}), \quad k = 1, \dots, K \quad (16)$$

The parameters  $\theta$  include all weight matrices and biases in the refinement network described below. The motivation for Partlets is: individual point features cannot reliably map to semantic part descriptions (e.g., a single point on a chair seat lacks context to predict "seat"), but aggregating features across all points in a part enables robust semantic grounding.

We initialize  $K$  learnable Partlet embeddings  $\{\mathbf{s}_k^{(0)} \in \mathbb{R}^{d_t}\}_{k=1}^K$  sampled from  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ , shared across all shapes but adapted per

instance through  $L$  refinement layers. At each layer  $\ell$ , each Partlet undergoes three operations:

**Partlet-to-Partlet Interaction:** Partlets interact to model part co-occurrence (e.g., chairs have seats and backs):

$$\mathbf{s}_k^{(\ell,1)} = \mathbf{s}_k^{(\ell-1)} + \text{SelfAttn}(\mathbf{s}_k^{(\ell-1)}, \{\mathbf{s}_{k'}\}_{k'=1}^K) \quad (17)$$

**Point-to-Partlet Aggregation:** Partlets gather shape-specific evidence from BiCo-fused point features:

$$\mathbf{s}_k^{(\ell,2)} = \mathbf{s}_k^{(\ell,1)} + \text{CrossAttn}(\mathbf{s}_k^{(\ell,1)}, \{\mathbf{h}_i\}_{i=1}^N) \quad (18)$$

**Non-linear Transformation:** Two-layer MLP with GELU:

$$\mathbf{s}_k^{(\ell)} = \mathbf{s}_k^{(\ell,2)} + \text{MLP}(\mathbf{s}_k^{(\ell,2)}) \quad (19)$$

After  $L$  layers, we obtain refined Partlet embeddings  $\mathbf{s}_k = \mathbf{s}_k^{(L)}$ .

**Mask Prediction.** Each Partlet predicts point membership via scaled dot-product:

$$m_{ki} = \frac{(\mathbf{W}_q \mathbf{s}_k)^T (\mathbf{W}_k \mathbf{h}_i)}{\sqrt{d_t}} \quad (20)$$

During training, we apply sigmoid activation; at inference, softmax across Partlets yields soft assignments.

**Partness Prediction.** Each Partlet predicts whether it represents an actual part:

$$\text{part}_k = \mathbf{w}_{\text{part}}^T \mathbf{s}_k + b_{\text{part}} \quad (21)$$

Higher values indicate active Partlets; lower values signify "no-part," enabling dynamic part count adaptation per shape.

**Semantic Alignment: Partlet  $\rightarrow$  Part Names.** Each Partlet's refined embedding  $\mathbf{s}_k$  serves directly as its prototype  $\mathbf{z}_k = \mathbf{s}_k \in \mathbb{R}^{d_t}$ . After normalization ( $\hat{\mathbf{z}}_k = \mathbf{z}_k / \|\mathbf{z}_k\|_2$ ), we compute cosine similarity with text embeddings  $\hat{\mathbf{t}}_a$ :

$$\text{sim}(k, a) = \hat{\mathbf{z}}_k \cdot \hat{\mathbf{t}}_a \quad (22)$$

Importantly, Partlet and text embeddings share the same semantic space ( $\mathbb{R}^{d_t}$ ) without intermediate projections - this alignment is driven by the text alignment loss (Section 3.2), enabling open-vocabulary matching at inference without retraining. However, part names are often ambiguous — *handle* could refer to a door handle, a mug handle, or a wheelchair handle, each with distinct geometry and function. To disambiguate, we use affordance-based descriptions (e.g., "the part of a door grasped to open it" for a door handle, "the horizontal surface of a chair where a person sits" for a chair seat) generated by Gemini 2.5 Flash [Comanici et al. 2025]. These are embedded with a sentence transformer model, such as MP-Net (all-mpnet-base-v2) [Song et al. 2020], for capturing long-form semantic affordances, particularly for similar part names.

### 3.2 Training: Partlet $\leftrightarrow$ Part Names

We establish correspondences between predicted Partlets and ground-truth parts via differentiable optimal transport. The cost matrix  $\mathbf{C} \in \mathbb{R}^{K \times A}$  combines mask overlap and semantic similarity:

$$C_{ka} = \mathcal{L}_{\text{mask}}^{(k,a)} + (1 - \text{sim}(\hat{\mathbf{z}}_k, \hat{\mathbf{t}}_a)) \quad (23)$$

where  $\mathcal{L}_{\text{mask}}^{(k,a)} = 1 - \text{Dice}(\sigma(\mathbf{m}_k), \mathbf{m}_a^{\text{gt}})$  (equal weighting  $\alpha = \beta = 1.0$ ).

Sinkhorn-Knopp iterations produce a soft assignment matrix  $\mathbf{P} \in [0, 1]^{K \times A}$ . Thresholding yields hard assignments  $\pi : \{1, \dots, K\} \rightarrow$

$\{1, \dots, A\} \cup \{\emptyset\}$ , where  $\pi(k) = a$  matches Partlet  $k$  to part  $a$  and  $\pi(k) = \emptyset$  indicates no match.

**Losses.** Let  $\mathcal{M} = \{k : \pi(k) \neq \emptyset\}$  denote matched Partlets. Our training objective combines several losses:

**Text Alignment Loss.** This loss is essential for open-vocabulary grounding. Without it, Partlet embeddings remain geometrically meaningful but semantically ambiguous. We apply InfoNCE contrastive loss to align matched Partlets with their text embeddings:

$$\mathcal{L}_{\text{text}} = \frac{1}{|\mathcal{M}|} \sum_{k \in \mathcal{M}} -\log \frac{\exp(\hat{\mathbf{z}}_k \cdot \hat{\mathbf{t}}_{\pi(k)} / \tau)}{\sum_{a=1}^A \exp(\hat{\mathbf{z}}_k \cdot \hat{\mathbf{t}}_a / \tau)} \quad (24)$$

with  $\tau = 0.07$ . Operating over Partlets rather than individual points makes this optimization tractable and stable.

**Mask and Partness Losses.** For matched Partlets, binary cross-entropy and Dice loss supervise masks:

$$\mathcal{L}_{\text{mask}} = \frac{1}{|\mathcal{M}|} \sum_{k \in \mathcal{M}} [\text{BCE}(\mathbf{m}_k, \mathbf{m}_{\pi(k)}^{\text{gt}}) + (1 - \text{Dice}(\sigma(\mathbf{m}_k), \mathbf{m}_{\pi(k)}^{\text{gt}}))] \quad (25)$$

Binary classification loss supervises partness, teaching Partlets to predict whether they are active (matched) or inactive (“no-part”):

$$\mathcal{L}_{\text{part}} = \frac{1}{K} \sum_{k=1}^K \text{BCE}(\text{part}_k, \mathbf{1}[\pi(k) \neq \emptyset]) \quad (26)$$

**Auxiliary Regularizers.** Coverage loss prevents over / under-segmentation by penalizing mask size disparities:

$$\mathcal{L}_{\text{cov}} = \frac{1}{|\mathcal{M}|} \sum_{k \in \mathcal{M}} \left| \frac{\sum_i \sigma(m_{ki}) - \sum_i m_{\pi(k)i}^{\text{gt}}}{N} \right| \quad (27)$$

Overlap loss enforces mutual exclusivity—each point should belong to at most one part:

$$\mathcal{L}_{\text{overlap}} = \frac{1}{N} \sum_{i=1}^N \left( \sum_{k=1}^K \sigma(m_{ki}) - 1 \right)^2 \quad (28)$$

**Global Alignment Loss.** Symmetric InfoNCE aligns the global shape representation with class-level text embeddings, providing object-level semantic context:

$$\mathcal{L}_{\text{global}} = \frac{1}{2} [\mathcal{L}_{\text{CE}}(\mathbf{S}, \mathbf{I}) + \mathcal{L}_{\text{CE}}(\mathbf{S}^T, \mathbf{I})] \quad (29)$$

where  $\mathbf{S} = \frac{1}{\tau} \hat{\mathbf{Z}}_{\text{global}} \hat{\mathbf{T}}_{\text{class}}^T$ .

**Total Loss.** The complete training objective as a function of model parameters  $\theta$  is:

$$\begin{aligned} \mathcal{L}_{\text{total}}(\theta) = & \lambda_{\text{mask}} \mathcal{L}_{\text{mask}}(\theta, \pi) + \lambda_{\text{part}} \mathcal{L}_{\text{part}}(\theta, \pi) \\ & + \lambda_{\text{text}} \mathcal{L}_{\text{text}}(\theta, \pi) + \lambda_{\text{cov}} \mathcal{L}_{\text{cov}}(\theta, \pi) + \lambda_{\text{ov}} \mathcal{L}_{\text{overlap}}(\theta, \pi) \\ & + \lambda_{\text{global}} \mathcal{L}_{\text{global}}(\theta) \end{aligned} \quad (30)$$

where  $\theta$  includes all learnable parameters (BiCo fusion weights, Partlet decoder layers, prediction heads),  $\pi$  is the assignment from Partlets to ground-truth parts, computed via Sinkhorn matching given the current model predictions. All hyperparameters are in the supplement.

### 3.3 Inference Modes and Use Cases

Our model supports three primary inference modes for different deployment scenarios.

**Mode 1: Closed-Vocabulary with Confidence Calibration.** We use ALIGN-Parts to scalably annotate large datasets from known categories (e.g., labeling millions of airplane meshes). This is the most practical scenario, enabling efficient 3D part segmentation and labeling with minimal oversight. We utilize this capability in building the TexParts dataset 4.7.

For training categories  $\mathcal{C}$ , we predict the object category via global shape-text alignment:

$$c^* = \arg \max_{c \in \mathcal{C}} \text{sim}(\mathbf{z}_{\text{global}}, \mathbf{t}_c) \quad (31)$$

then filter Partlets by partness score ( $\sigma(\text{part}_k) > 0.5$ ) to obtain active set  $\mathcal{K}_{\text{active}}$ . We construct cost matrix  $\mathbf{C} \in \mathbb{R}^{|\mathcal{K}_{\text{active}}| \times |\mathcal{L}_{c^*}|}$ :

$$C_{ka} = 1 - \text{sim}(\hat{\mathbf{z}}_k, \hat{\mathbf{t}}_a) \quad (32)$$

At inference, we use the Jonker-Volgenant algorithm [Crouse 2016] for exact optimal assignment, which is more efficient than Sinkhorn since gradients are not required.

**Mahalanobis Score Estimation.** The Mahalanobis confidence (Eq. 35) requires class-conditional statistics (mean and covariance) that are estimated from the training set.

After training, we perform a single forward pass over the entire training dataset. For every partlet  $k$  that is successfully matched to a ground-truth part label  $\ell$  (i.e.,  $\pi(k) = \ell$ ), we extract its prototype embedding  $\mathbf{z}_k$ .

We then compute the empirical mean  $\boldsymbol{\mu}_\ell$  for each part label  $\ell$  in our known training vocabulary  $\mathcal{C}$ :

$$\boldsymbol{\mu}_\ell = \mathbb{E}[\mathbf{z}_k | \pi(k) = \ell] \quad (33)$$

For robustness, we compute a single, shared covariance matrix  $\Sigma$  by pooling the embeddings from all part classes:

$$\Sigma = \text{Cov}(\{\mathbf{z}_k\}_{\forall k, \ell \text{ s.t. } \pi(k) = \ell}) \quad (34)$$

We apply regularization (e.g., adding a small value  $\epsilon \mathbf{I}$  to the diagonal) before computing the inverse  $\Sigma^{-1}$  to ensure numerical stability. These pre-computed  $\boldsymbol{\mu}_\ell$  and  $\Sigma^{-1}$  are stored and used at inference time to compute the following Mahalanobis-distance based confidence score:

$$\text{conf}_{\text{maha}}(k) = \exp(-(\mathbf{z}_k - \boldsymbol{\mu}_{a_k^*})^T \Sigma_{a_k^*}^{-1} (\mathbf{z}_k - \boldsymbol{\mu}_{a_k^*})) \quad (35)$$

where  $\boldsymbol{\mu}_{a_k^*}$  and  $\Sigma_{a_k^*}$  are estimated from training embeddings. Predictions with  $\text{conf}_{\text{maha}} \geq 0.8$  are auto-accepted; lower-confidence predictions are flagged for human verification, dramatically reducing annotation cost.

**Fused Confidence Formulation.** The final confidence score  $\text{conf}(k)$  for a matched query  $k$  is a fusion of the softmax confidence ( $\text{conf}_{\text{soft}}$ ) and the Mahalanobis confidence ( $\text{conf}_{\text{maha}}$ ). We combine them as follows:

$$\text{conf}(k) = \alpha \cdot \text{conf}_{\text{soft}}(k) + (1 - \alpha) \cdot \sigma(\beta \cdot (\text{conf}_{\text{maha}}(k) - 0.5)) \quad (36)$$

where  $\sigma(\cdot)$  is the sigmoid function.

- $\text{conf}_{\text{soft}}(k)$  is the temperature-calibrated softmax score (Eq. 38).
- $\text{conf}_{\text{maha}}(k)$  is the Mahalanobis confidence (Eq. 35).



- $\alpha$  and  $\beta$  are hyperparameters that balance the two scores. We set  $\alpha = 0.5$  and  $\beta = 1.0$  based on calibration on a held-out validation set.

Annotations with  $\text{conf}(k) < \tau_{\text{conf}}$  (where  $\tau_{\text{conf}} = 0.5$ ) are flagged as low-confidence and routed to a human annotator for manual review.

**Mode 2: Open-Vocabulary Grounding.** For novel categories, users provide candidate part descriptions  $\{\mathbf{t}_a\}_{a=1}^A$  (or generate via LLM from image/hint). After filtering inactive Partlets, each active Partlet matches to the best description:

$$a_k^* = \arg \max_{a \in \{1, \dots, A\}} \text{sim}(\hat{\mathbf{z}}_k, \hat{\mathbf{t}}_a) \quad (37)$$

Confidence is computed via temperature-calibrated softmax:

$$\text{conf}_{\text{soft}}(k) = \max_a \frac{\exp(\hat{\mathbf{z}}_k \cdot \hat{\mathbf{t}}_a / \tau)}{\sum_{a'=1}^A \exp(\hat{\mathbf{z}}_k \cdot \hat{\mathbf{t}}_{a'} / \tau)} \quad (38)$$

with  $\tau = 0.07$ . This is less calibrated than the Mahalanobis distance (Mode 1) due to the lack of training statistics for novel categories.

**Mode 3: Text-Conditioned Part Retrieval.** For compatibility with Find3D [Ma et al. 2024], we retrieve a single part for query  $\mathbf{t}_q$ :

$$k^* = \arg \max_{k \in \mathcal{K}_{\text{active}}} \text{sim}(\hat{\mathbf{z}}_k, \hat{\mathbf{t}}_q) \quad (39)$$

returning mask  $\mathbf{m}_{k^*}$ . This mode is primarily for benchmark comparison.

**Point Label Assignment.** Points are assigned to the highest-scoring Partlet’s label:

$$\text{label}_i = a_{k^*}^* \quad \text{where} \quad k^* = \arg \max_{k \in \mathcal{K}_{\text{active}}} \sigma(m_{ki}) \quad (40)$$

Points with  $\max_k \sigma(m_{ki}) < 0.5$  remain unlabeled.

### 3.4 Implementation Details

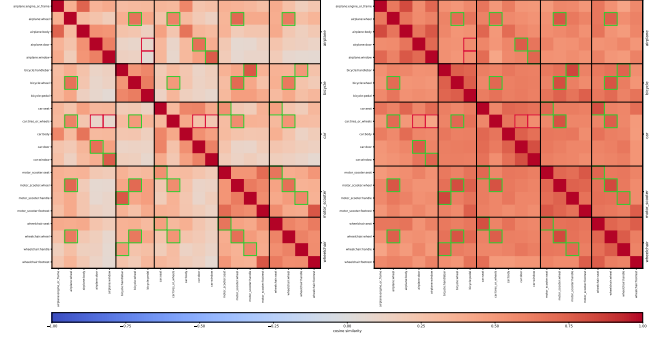
**Training Setup.** ALIGN-Parts is trained on 3 NVIDIA A6000 GPUs for 2 days (batch size 16). **Due to academic compute constraints, we sample 10k points per shape (vs. 100k in PartField)** while maintaining strong performance. Models are normalized to  $[-1, 1]^3$  during training.

**Architecture.** The model has 34M parameters total: 5.7M for feature Fusion, 26.8M for Partlets, and 1.5M for the global classifier. Feature interactions (Partlet-to-points and Partlet-to-Partlet) use 3 transformer [Vaswani et al. 2017] blocks with multi-head cross-attention, LayerNorm, residual connections, and feedforward layers. The BiCo fusion employs sparse 16-NN attention with 3D relative positional bias, computed via a learned MLP over Fourier-encoded ( $F=6$  frequencies) displacement vectors, which provides geometric context while maintaining  $\mathcal{O}(Nk)$  complexity. We also note that for calculating the runtime, we do not include data preprocessing time, as it varies depending on parallelization and system capabilities.

We set the number of Partlets to 32, as this value provides a reasonable estimate for the typical number of semantic parts found in most objects in our unified dataset. This choice is further validated by analyzing the statistics of part counts across the full dataset,

which confirms that 32 accommodates the majority of objects without excessive over-segmentation or loss of fine granularity.

**Optimization.** Loss weights:  $\lambda_{\text{mask}} = 1.0$ ,  $\lambda_{\text{part}} = 0.5$ ,  $\lambda_{\text{text}} = 1.0$ ,  $\lambda_{\text{cov}} = 0.5$ ,  $\lambda_{\text{overlap}} = 0.1$ ,  $\lambda_{\text{global}} = 1.0$ . We use AdamW with an initial learning rate of  $3\text{e-}4$  and cosine annealing to a minimum of  $5\text{e-}6$ .



**Figure 4: Pairwise cosine similarity heatmaps between text embeddings for MPNet (left) and SigLIP (right). (Zoom in for labels)**

**Choice of Text Encoder** We adopt MPNet [Song et al. 2020] as our text encoder for part descriptions rather than CLIP / SigLIP due to its superior structure-preserving properties for sentence-level embeddings. SigLIP, optimized for short image captions (e.g., "A photo of a dog"), exhibits pathologically high cosine similarities across semantically distinct part descriptions generated by Gemini, undermining the discriminative structure necessary for partlet-based learning. In Fig. 4 for instance, MPNet correctly assigns high similarities ( $>0.8$ ) to functionally equivalent parts across classes - such as wheels (airplane, car, bicycle, wheelchair), doors (airplane, car), and handles (scooter, bicycle, wheelchair), while maintaining low similarities ( $<0.3$ ) between parts with different affordances, such as tires vs. doors/windows or pedals vs. airplane components. In contrast, SigLIP assigns uniformly high similarities to both sets, collapsing the semantic space and preventing our partlets from learning meaningful text-conditioned part alignment during training.

**Why affordance descriptions?** A key motivation for incorporating affordance information into part annotations is rooted in the cognitive science understanding that humans interpret and define object parts not just by geometry, but by their function, context, and description. Short or generic part names (e.g., "leg", "handle") are often ambiguous across different objects, lacking any semantic detail regarding the role or meaning of a part within a specific context. For example, "legs" fulfill distinct structural functions and take on different forms for chairs, tables, or sofas, a distinction that arises from their object-specific affordances. Prior work demonstrates that affordance-based cues and descriptive information guide human part recognition, reducing label ambiguity and supporting more robust reasoning and communication. Thus, by situating part annotations within functional and contextual descriptions, our approach enables higher-quality, less ambiguous labeling, consistent with

cognitive models of human object understanding.

**Datasets.** We train on 40,982 shapes from three datasets: 3DCoMPaT++ (8,627), PartNet (32,141), and Find3D (124). All use fine-grained part labels. For evaluation, we hold out 206 shapes: 126 objects (42 categories) from 3DCoMPaT++, 72 objects (24 categories) from PartNet, and 8 novel objects (8 categories) from Find3D.

## 4 Experiments

In this section, we discuss our methodology for unifying 3D part benchmarks with inconsistent part naming conventions for the same object classes (Section 4.1), introduce our baselines and metrics for the named 3D part segmentation task (Section 4.2), and show qualitative and quantitative comparison of ALIGN-Parts with related baselines (Section 4.3). We ablate different components and inference modes of ALIGN-Parts in Section 4.4 and Section 4.5. We end the section with 2 applications of our method in Section 4.6 and Section 4.7.

### 4.1 3D Part Annotation Alignment & Vocabulary Compression

Our annotation alignment and vocabulary compression system employs a two-stage pipeline that combines MPNet embeddings for candidate generation and Gemini LLM verification to reject spurious matches as well as identify and merge duplicate classes and parts across a unified 3D object taxonomy. For confirmed matches, the system successfully identifies semantically equivalent entities with high MPNet similarity scores that Gemini validates as identical: for example, "laptop\_computer" and "laptop" (similarity: 0.944) are merged because Gemini recognizes that "both candidate names refer to the exact same physical device... consistently define it as a portable personal computer designed for mobile use". Similarly, *microwave\_oven* (3DCoMPaT) and *microwave* (PartNet) with similarity 0.902 are merged after Gemini confirmed they "describe the same kitchen appliance." Within the "microwave\_oven" class, "door\_glass" and "glass" (similarity: 0.865) are unified because Gemini concludes "both descriptions refer to the transparent panel integrated into the door... that allows viewing food and contains radiation. The secondary part name 'glass' is a concise reference to the 'door\_glass'. Similarly, "bed\_footboard" and "footboard" (similarity: 0.953) are merged as Gemini states, "bed\_footboard" is a more explicit naming of 'footboard', and their descriptions are semantically identical, describing a panel at the foot of the bed opposite the headboard." For rejected pairs, the system correctly distinguishes semantically distinct parts despite high embedding similarity: "car\_front\_bumper" and "car\_rear\_bumper" (similarity: 0.879) are kept separate because Gemini determines "while both parts are bumpers with the same protective function, their specified locations (front vs. rear) make them distinct semantic parts for a 3D car object," and within the "chair" class, "back\_frame\_horizontal\_rod" and "back\_frame\_vertical\_rod" (similarity: 0.943) remain separate because Gemini explains "the parts are distinct based on their orientation within the back frame: one is explicitly described as a 'horizontal rod' providing reinforcement for the backrest, while the other is a 'vertical rod' providing structural support." The compressed vocabulary output maintains canonical names (choosing

more verbose/descriptive variants), aggregates part counts across merged entities, and produces a mapping log that records every alias resolution for downstream lookup when legacy names are encountered during inference. This compressed vocabulary enables training with unified part semantics.

### 4.2 Comparison on Named 3D Part Segmentation

**Baselines.** While no prior work addresses *named* 3D part segmentation end-to-end, we compare against two recent baselines: PartField [Liu et al. 2025] and Find3D [Ma et al. 2024]. PartField achieves state-of-the-art class-agnostic segmentation but cannot name parts. It clusters a learned feature field (requiring ground-truth part count  $K$ ). Find3D matches per-point features to SigLIP embeddings of *provided* part queries, needing the list of ground-truth part names as input.

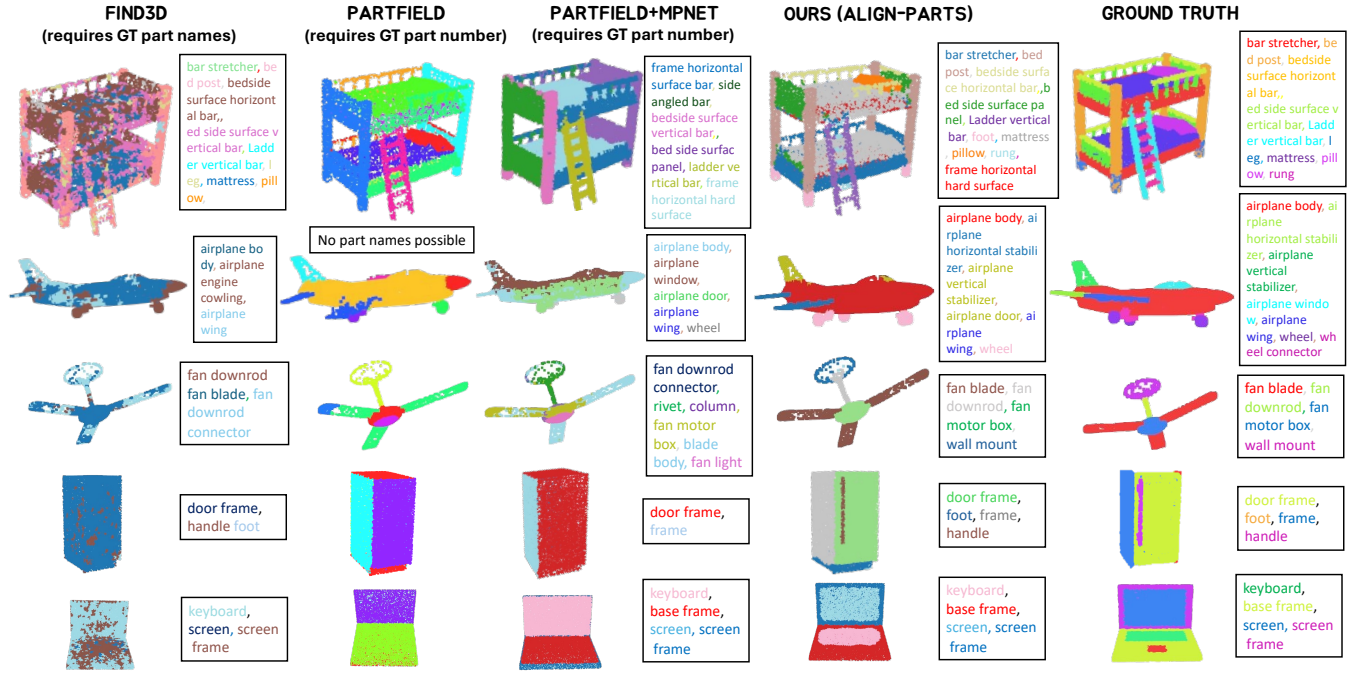
**PartField+MPNet (Our Strong Baseline; w/o partlets).** We extend PartField with semantic alignment: a linear head maps PartField features to a shared space with MPNet text embeddings, trained via InfoNCE loss. However, it still requires predicting  $K$  via an auxiliary classifier, making it brittle to over/under-segmentation from  $K$  prediction errors. Further architecture details are provided in Section A.1.

**Metrics.** We evaluate our method using three complementary metrics that progressively incorporate semantic label correctness. (1) Class-agnostic mIoU: Following prior work [Liu et al. 2025; Yang et al. 2024], for each ground-truth part, we compute the maximum IoU across all predicted segments and average these values, ignoring semantic labels entirely - this captures pure geometric segmentation quality. (2) Label-Aware mIoU (strict, LA-mIoU): For each ground-truth part, we identify the predicted segment with the highest geometric overlap (as in class-agnostic mIoU), then assign credit only if its semantic label exactly matches the ground truth; otherwise, the part contributes 0.0 - this measures joint geometry-semantic accuracy with strict label matching. (3) Relaxed Label-Aware mIoU (rLA-mIoU): Identical segment selection to strict LA-mIoU, but instead of binary label matching, we weight the IoU by the cosine similarity between MPNet text embeddings of predicted and ground-truth labels, giving partial credit to semantically related predictions (e.g., "screen" vs. "monitor") - this captures semantic near-misses that strict matching penalizes. By construction, we have class-agnostic mIoU  $\geq$  rLA-mIoU  $\geq$  LA-mIoU (strict), where equality holds only when all predictions have perfect label agreement. The gap between class-agnostic and label-aware metrics reveals semantic prediction errors, while the gap between strict and relaxed variants quantifies label confusion on semantically similar parts. Correlation plots (Fig 6 and 7) empirically validate that relaxed scoring recovers significant semantic credit on near-miss predictions.

### 4.3 Results

In Figures 5 and 8 and Table 2, we show that ALIGN-Parts comprehensively outperforms all baselines on both class-agnostic segmentation (mIoU) and named part segmentation (LA-mIoU). On average mIoU, we outperform PartField by 15.8%, whereas on LA-mIoU and rLA-mIoU, we improve over PartField+MPNet by 58.8% and 43.8%,





**Figure 5: Qualitative Results.** ALIGN-Parts segments and names 3D parts robustly in a single feed-forward pass (rightmost column). Find3D [Ma et al. 2024] (first column) fails despite ground-truth part names, unable to segment the laptop in Figure 5 (bottom row). PartField [Liu et al. 2025] (second column) also fails: it requires ground-truth part counts for clustering, missegments bed bunks (top row), and misses refrigerator handles (second-to-last row). Our strong baseline without Partlets (third column) exhibits similar errors. In contrast, ALIGN-Parts correctly segments tiny parts, such as handles, and groups semantically similar instances (e.g., all ceiling fan blades into a single cluster).

**Table 2: Evaluation of ALIGN-Parts and related baselines on our test set.** In addition to the usual mIoU metric for evaluating class-agnostic part segmentation, we introduce 2 metrics more suited for our named part-segmentation task, namely, label-aware mIoU (LA-mIoU) and relaxed label-aware mIoU (rLA-mIoU).

Variant	3DComPaT			Find3D			PartNet			Average		
	mIoU↑	LA-mIoU↑	rLA-mIoU↑	mIoU↑	LA-mIoU↑	rLA-mIoU↑	mIoU↑	LA-mIoU↑	rLA-mIoU↑	mIoU↑	LA-mIoU↑	rLA-mIoU↑
PartField	0.371	n/a	n/a	0.662	n/a	n/a	0.521	n/a	n/a	0.518	n/a	n/a
Find3D	0.239	0.072	0.178	0.379	0.232	0.324	0.354	0.204	0.298	0.324	0.169	0.267
PartField+MPNet	0.316	0.185	0.259	0.590	0.137	0.451	0.446	0.276	0.394	0.451	0.199	0.368
ALIGN-Parts	0.453	0.268	0.391	0.595	0.133	0.466	0.753	0.546	0.729	0.600	0.316	0.529

respectively, while being 100× faster because we do not require running a K-means clustering algorithm.

Qualitatively, our baselines show several weaknesses. PartField often fragments instances of the same part into multiple segments after clustering, contradicting human labeling conventions in 3DComPaT++ and PartNet. This likely stems from its use of SAM [Kirillov et al. 2023] to extract unlabeled parts from 2D renderings. In contrast, ALIGN-Parts, trained on human annotations, correctly groups instances: e.g., all four bed posts of a double bed and all three wheels of an airplane are identified as single parts (Figure 8).

Find3D learns per-point semantic vectors without considering shape geometry, resulting in noisy and overlapping segmentations. PartField+MPNet’s reliance on predicted cluster counts leads to

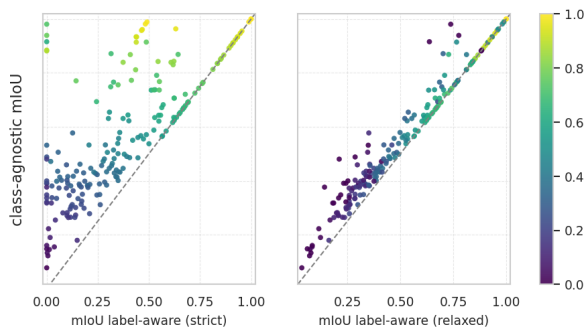
under-segmentation, as fine parts such as the refrigerator handle and laptop screen frame are missed (Figure 8). Both Find3D and PartField fail to segment even simple structures, such as fan blades, whereas ALIGN-Parts accurately segments and annotates fine-grained parts in complex shapes.

**Runtime Comparison.** In terms of runtime, our method compares favorably against all baselines. Find3D runs in  $\approx 0.25s$ , whereas both PartField and PartField+MPNet require  $\approx 4s$ , where the majority of the runtime is consumed by K-means clustering. ALIGN-Parts, being a one-shot feedforward method, needs only  $\approx 0.05s$  (barring feature pre-processing) to produce labelled 3D parts.

**Fine-Part Localization** Despite using only 10k sampled points, ALIGN-Parts segments fine parts like the *screw* of scissors (Fig 8) - structures that PartField, trained with 100k points, cannot localize.

**Table 3: Ablation study on 3DComPaT, Find3D, and PartNet. We report mIoU, label-aware mIoU (LA-mIoU), and relaxed label-aware mIoU (rLA-mIoU).**

Variant	3DComPaT			Find3D			PartNet			Average		
	mIoU↑	LA-mIoU↑	rLA-mIoU↑	mIoU↑	LA-mIoU↑	rLA-mIoU↑	mIoU↑	LA-mIoU↑	rLA-mIoU↑	mIoU↑	LA-mIoU↑	rLA-mIoU↑
Base Model	0.228	0.021	0.141	0.374	0.041	0.239	0.335	0.027	0.201	0.312	0.030	0.194
No $\mathcal{L}_{cov}$ , $\mathcal{L}_{ov}$ , $\mathcal{L}_{txt}$	0.233	0.019	0.140	0.382	0.003	0.198	0.357	0.041	0.223	0.324	0.021	0.187
No $\mathcal{L}_{cov}$ , $\mathcal{L}_{ov}$	0.422	0.193	0.338	0.499	0.110	0.384	0.664	0.414	0.609	0.528	0.239	0.443
Geo Input Only	0.224	0.014	0.134	0.384	0.027	0.187	0.332	0.040	0.210	0.313	0.027	0.177
Feature Concat	0.221	0.017	0.134	0.367	0.058	0.239	0.317	0.033	0.194	0.302	0.036	0.189
PartField+MPNet	0.316	0.185	0.259	0.590	0.137	0.451	0.446	0.276	0.394	0.451	0.199	0.368
ALIGN-Parts	0.453	0.268	0.391	0.595	0.133	0.466	0.753	0.546	0.729	0.600	0.316	0.529

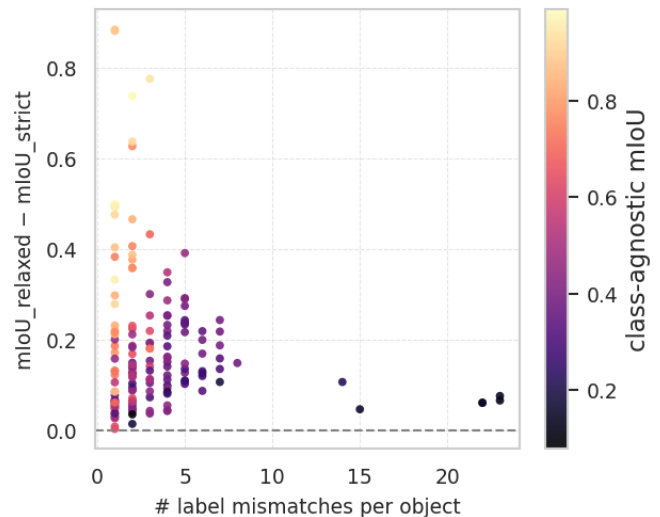
**Figure 6: Proposed Metrics Correlation Analysis.** Correlation analysis between our proposed label-aware mIoU metrics (strict/relaxed) and class-agnostic mIoU, computed on segmentation results from our ALIGN-Parts model. The strict label-aware metric (left) shows moderate agreement with class-agnostic mIoU (Pearson  $r = 0.739$ , Spearman  $\rho = 0.730$ ,  $N = 206$ ), while the relaxed variant (right) demonstrates near-perfect correlation (Pearson  $r = 0.978$ , Spearman  $\rho = 0.974$ ,  $N = 206$ ). These findings indicate that our model achieves strong semantic and quantitative consistency, further supporting the use of the relaxed metric as a robust evaluation protocol for semantic 3D part segmentation.

#### 4.4 Ablations

Table 3 and Figure 9 evaluate ALIGN-Parts’s design choices, showing quantitative and qualitative improvements from each component.

**Baseline Comparisons.** Using only PartField geometric features, our base model ( $\mathcal{L}_{mask}$ ,  $\mathcal{L}_{part}$ ,  $\mathcal{L}_{global}$ ) performs slightly better than naive concatenation of PartField and DINOv2 features, indicating that simple multi-modal fusion can be counterproductive. A variant using raw geometry without PartField shows no improvement, confirming that learned geometric features are essential.

**Progressive Component Addition.** Adding DINOv2 appearance features yields modest gains. Incorporating the InfoNCE text alignment loss ( $\mathcal{L}_{txt}$ ) significantly improves both LA-mIoU and rLA-mIoU; this variant is the only one, besides our full model, that correctly segments fine-grained parts, such as plants in vases or shelves in storage furniture. Finally, adding auxiliary regularizers ( $\mathcal{L}_{cov}$ ,  $\mathcal{L}_{overlap}$ ) yields the complete ALIGN-Parts model, which achieves peak performance both qualitatively and quantitatively.

**Figure 7: Label Mismatches and Metric Robustness.** Distribution of test objects according to the number of label mismatches (x-axis) and the resulting increase in mIoU from strict to relaxed label-aware matching (y-axis), colored by class-agnostic mIoU. Most objects with low mismatches exhibit small or moderate increases in mIoU, while objects with higher mismatch counts still do not show large outliers in metric difference, supporting the robustness of our evaluation protocol. The absence of extreme discrepancies suggests that the relaxed metric yields a stable and meaningful improvement, even for challenging cases, confirming its reliability for assessing semantic part segmentation on our test set.**Figure 8: ALIGN-Parts correctly segments the tiny screw despite training with sparser points (10k vs. 100k).**

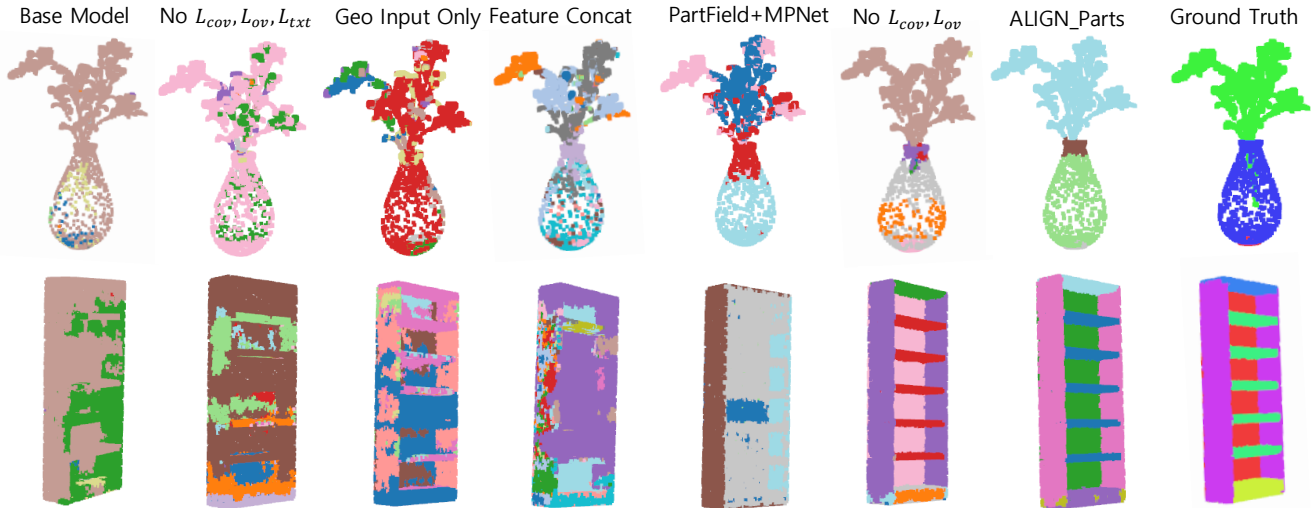


Figure 9: Ablation (Qualitative results). Results improve from left to right as components are added sequentially, with ground truth in the final column. The first five columns result in significant misalignments and segmentation leakage. Neither geometric-only features nor naive DINO concatenation improves performance. Major gains arise from Partlets (see sixth and seventh columns) and the coverage loss, which refine fine details and are consistent with quantitative metrics in Table 3.

#### 4.5 Inference-time Ablations

We evaluate two additional inference modes of our ALIGN-Parts model, extending beyond the primary dynamic part activation approach to better understand the contributions of part cardinality and label information in our segmentation pipeline. The first alternative mode, which we term the *clustering+part number* setting, completely forgoes the use of any part vocabularies or text labels during inference. Instead, it relies solely on the fused geometric and appearance features output by the model, upon which we run k-means clustering to produce purely class-agnostic instance clusters. This setup rigorously probes the ability of the learned feature embeddings, untethered to semantic labels, to support coherent part decompositions across diverse objects, essentially isolating the impact of visual and geometric cues alone. The second mode, called *+Part number*, examines whether providing the model with the exact ground-truth part count for each input shape improves segmentation quality compared to the default setting, where the model dynamically infers the number of parts to activate. After producing all candidate partlet masks and calculating their partness scores, this mode ranks the partlets by a saliency score, which is a composite measure combining the confidence that a partlet corresponds to an actual part (i.e., partness) and the average mask coverage over the point cloud (mean mask probability mass over points). From this ranking, the top  $M$  partlets are retained, where  $M$  is the true number of parts for the target shape, and every point in the shape is assigned the best matching mask among these selected partlets to yield a hard  $K$ -way partition. These inference ablation modes and their qualitative and quantitative outcomes are detailed and visualized in Fig 10 and Tab 4, demonstrating that ALIGN-Parts is able to robustly estimate accurate part cardinality and segmentation even without explicit part label or count guidance, and that the fused multimodal features alone provide meaningful cues towards

coherent part delineation. This analysis not only highlights the flexibility and robustness of ALIGN-Parts at inference time, but also emphasizes the benefits of its design choices in learning and leveraging rich feature representations supporting both semantic and instance-aware part segmentation.

#### 4.6 3D Shape Co-Segmentation and Part Label Transfer

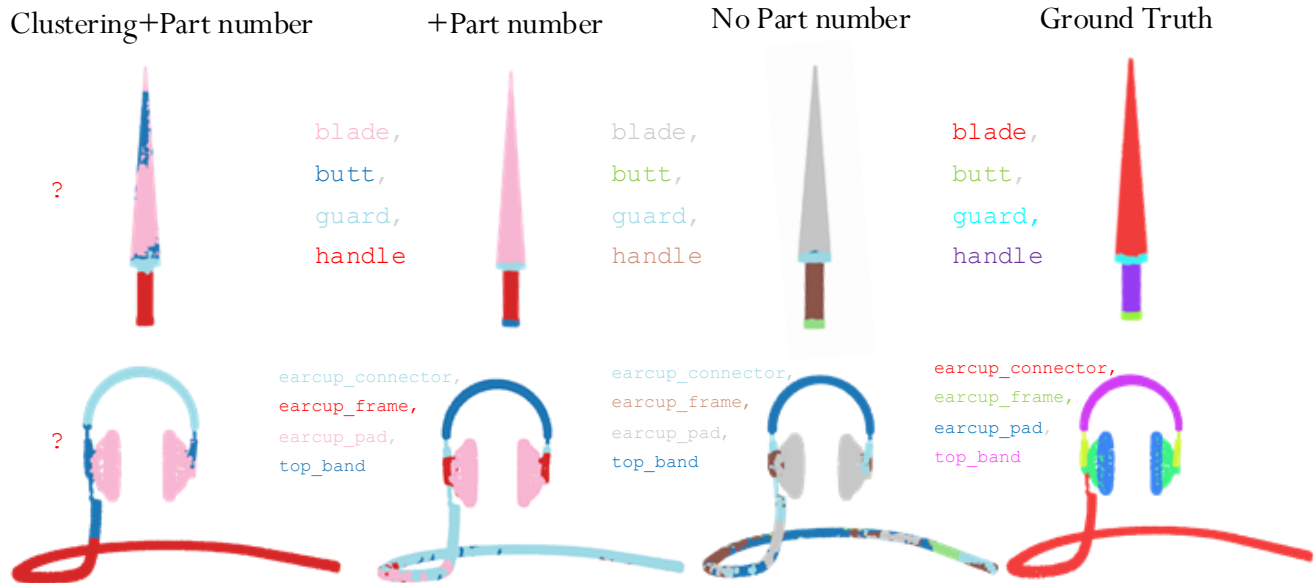
Fig. 11 shows results and analysis of 3D Shape Co-Segmentation using ALIGN-Parts (and the BiCo features) as compared to PartField. A classical approach to 3D part segmentation operates in a co-segmentation setting, where multiple shapes from the same category are jointly analyzed to establish consistent part correspondence. Prior methods, including PartField, employ what we call *dependent clustering*: they first segment a source shape using feature clustering, then initialize k-means clustering on a target shape using the source cluster centroids, implicitly enforcing part correspondence. While this strategy can succeed on geometrically similar shapes, it proves fragile when target shapes exhibit different part counts or topologies. For example, in Figure 11, dependent clustering on a moderately similar target chair (Target A) produces reasonable results, but fails dramatically on targets with substantial part variation (Target B), causing geometric boundaries to blur (e.g., the backrest merging incorrectly with the seat).

An alternative, *independent clustering* approach segments each target shape autonomously and then matches clusters post hoc by comparing source and target cluster centers. As shown in Target C, this mode is more robust to topological differences, though it forgoes any direct geometric correspondence to the source.

In contrast to both clustering-based paradigms, our proposed ALIGN-Parts adopts a fully *feedforward*, discriminative approach that predicts part segmentation masks and semantic labels jointly,

**Table 4: Evaluation of different inference modes of ALIGN-Parts on our test set, using mean IoU (mIoU), label-aware mIoU (LA-mIoU), and relaxed label-aware mIoU (rLA-mIoU). Providing additional ground-truth part count information only slightly improves the model’s performance on rLA-mIoU showing that ALIGN-Parts often estimates accurate part cardinality based on just the input geometric and appearance features of a 3D shape.**

Variant	3DCoMPaT (126)			Find3D (8)			PartNet (72)			Average		
	mIoU↑	LA-mIoU↑	rLA-mIoU↑	mIoU↑	LA-mIoU↑	rLA-mIoU↑	mIoU↑	LA-mIoU↑	rLA-mIoU↑	mIoU↑	LA-mIoU↑	rLA-mIoU↑
Clustering + Part Number	0.370	n/a	n/a	0.528	n/a	n/a	0.537	n/a	n/a	0.478	n/a	n/a
+Part Number	0.452	0.268	0.389	0.625	0.138	0.473	0.757	0.559	0.737	0.611	0.322	0.533
No Part Number	0.453	0.268	0.391	0.595	0.133	0.466	0.753	0.546	0.729	0.600	0.316	0.529



**Figure 10: Inference-time Ablations.** We assess two additional inference modes in our ALIGN-Parts model. The *Clustering+Part number* mode disregards text vocabularies and applies k-means to the fused geometric and appearance features, generating class-agnostic clusters and probing how well raw feature representations alone support meaningful part segmentation. The *+Part number* mode uses the true part count for each shape, activating exactly  $M$  partlets with the highest saliency scores - computed by combining partness confidence and average mask coverage - and assigning every point to its best matching mask; this tests whether supplying ground-truth cardinality adds value compared to dynamic, data-driven part discovery. Both setups are compared to the default dynamic activation (No Part number) and the ground truth, highlighting that ALIGN-Parts robustly estimates part cardinality and produces accurate segmentations even without explicit part label or count supervision.

without requiring source shape initialization or explicit co-segmentation. This design eliminates brittleness to part count variation and geometric mismatches, enabling robust generalization across shapes with diverse part structures and semantics. As demonstrated in Figure 11 (middle and right panels), ALIGN-Parts consistently produces accurate, semantically grounded part segmentations regardless of target shape complexity.

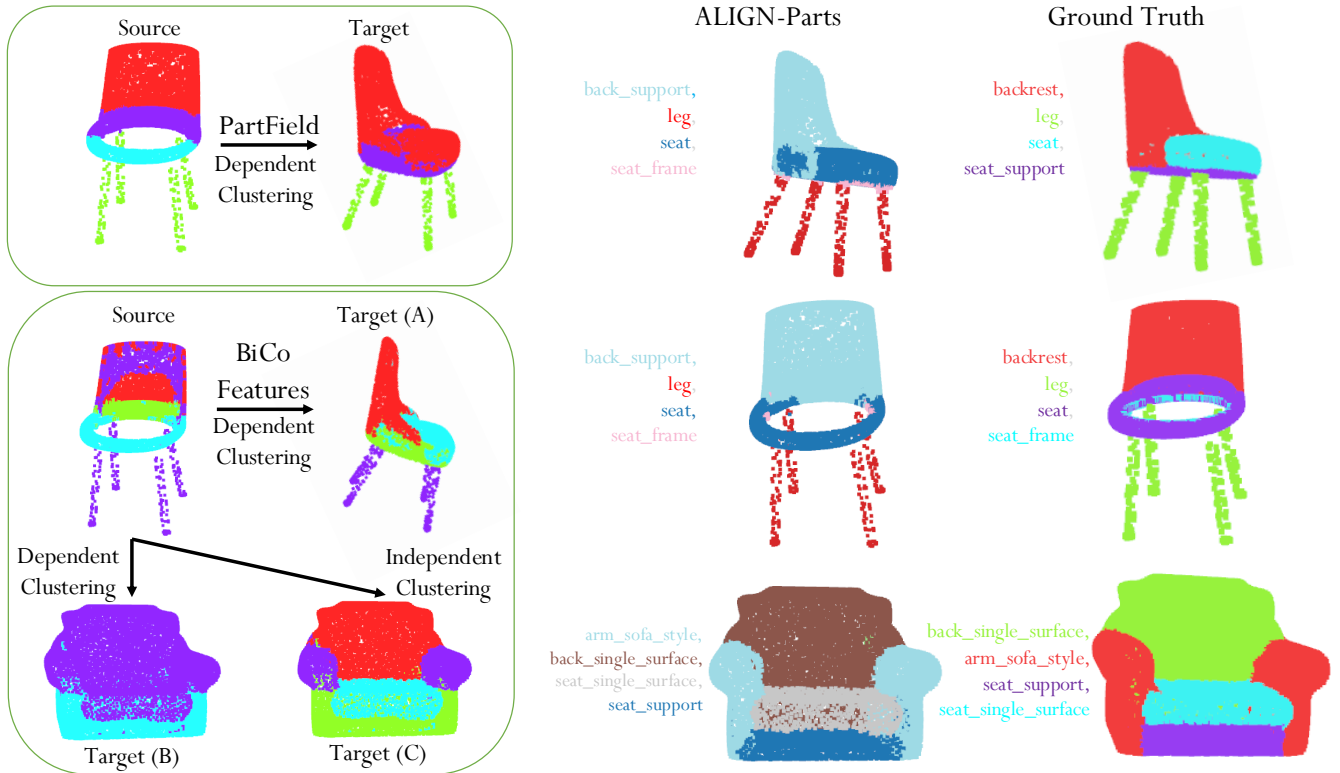
#### 4.7 TexParts Dataset

A central aim of our approach is to enable the construction of a high-quality 3D part annotation dataset with minimal manual intervention, ensuring both unified and comprehensive part labeling at scale. For this purpose, we select the TexVerse dataset

as our unannotated 3D source corpus, leveraging its exceptional quality, high-resolution textures, and extensive diversity of 3D assets [Zhang et al. 2025]. TexVerse consists of over 850,000 unique 3D models with physically based rendering (PBR) materials and rich metadata, making it an ideal foundation for large-scale part segmentation.

Our pipeline begins with the automated filtering of TexVerse models: using Gemini-Flash LLM, we combine thumbnail images and other metadata to preselect high-quality objects and exclude inadequate or malformed models. Next, we apply our ALIGN-Parts model and save, for each shape, its predicted part masks, part names, and both semantic and segmentation confidence scores. To prioritize downstream annotation effort, we sort objects by their average





**Figure 11: 3D Shape Co-Segmentation Analysis.** *Left: clustering-based co-segmentation.* Prior methods such as PartField perform *dependent clustering* by first segmenting a source shape via feature clustering and then using the resulting cluster means to initialize K-means on a target shape, implicitly enforcing part correspondence; this can break when the target has a different part count or geometry, causing errors such as the red backrest region bleeding into the seat on the target chair. Using the same dependent co-segmentation strategy with our fused BiCo features (BiCo Feature Dependent Clustering) yields improved transfers on moderately similar targets (Target A), but performance degrades on more challenging targets with greater variation in part structure (Target B). As an alternative, we apply independent clustering to Target C, where the target is segmented with source initialization and clusters are matched post hoc by comparing source and target cluster centers, which proves more reliable for difficult co-segmentation cases. *Middle and right: feedforward ALIGN-Parts.* In contrast to all clustering-based variants, the proposed feedforward ALIGN-Parts model (middle) directly predicts part segmentation and names, achieving robust results across shapes with differing part counts and topologies, and eliminating any dependence on source shapes or explicit co-segmentation.

confidence score (in descending order) so that annotators see the most reliable candidates first. Selected objects are then routed to human annotators for validation and correction.

During annotation, annotators use several aids: a part name prompting tool for searching or extending the active part vocabulary, and (optionally) the ability to reference unlabeled geometric part masks generated by PartField. Our annotation process is explicitly bilevel; Phase One focuses on validating and making minor edits to ALIGN-Parts predictions, while Phase Two addresses new or missing parts that require more substantive manual annotation. By the time of submitting this work, the first phase has covered approximately 8,000 objects, comprising around 14,000 unique part categories. Examples from the current dataset are shown in Figure 12.

A key observation from our annotation workflow is the dramatic reduction in manual effort enabled by our methodology: annotating 3D objects from scratch typically takes anywhere from 15 to 25 minutes per shape, while our model-assisted pipeline reduces annotation time to just 3 to 5 minutes on average - a time saving of approximately 5–8× without sacrificing annotation quality.

Importantly, and in clear contrast to recent approaches that keep their Objaverse-derived part annotations closed-source, we commit to releasing TexParts as a public resource upon publication, with the aim of advancing large-scale open research in semantic 3D part understanding.

## 5 Limitations and Future Work

The primary limitations of our work stem from the relatively restricted set of objects and parts on which ALIGN-Parts has been



**Figure 12: TexParts Dataset. We demonstrate human-in-the-loop annotation of Texverse [Zhang et al. 2025] using ALIGN-Parts, enabling scalable dense 3D part segmentation.**

trained, compared to the vast (though finite) variety of parts that occur in the real world. This gap is largely due to the scarcity of large-scale 3D datasets with dense part annotations and a unified, operational definition of what constitutes a “part”. In effect, this creates a chicken-and-egg problem: ALIGN-Parts was designed to enable robust 3D part annotation at scale, yet the robustness and coverage of the model itself are constrained by the limited annotated data available for training.

Future work will focus on mitigating this dependency by exploring self-supervised or weakly supervised formulations and by incorporating stronger 3D priors, for example from generative models or skeletal/medial representations. Another important direction is to reduce the current reliance on frozen PartField features by enabling full end-to-end training of the geometric feature extractor, which was not pursued here primarily due to computational constraints rather than methodological ones. Despite these limitations, our framework is immediately usable by organizations and research labs with abundant compute and proprietary 3D assets, who can scale ALIGN-Parts to richer, closed-source datasets and drive progress towards truly large-scale 3D scene understanding at the part level.

## 6 Discussion

ALIGN-Parts reframes *semantic* 3D part segmentation as a set alignment problem, where Partlets trained via optimal transport jointly

learn geometry and semantics without part count supervision. Unlike Find3D’s per-point alignments or PartField’s brittle clustering, our end-to-end matching produces fast, coherent, named parts directly. LLM-generated affordance descriptions (“door handle is grasped to open a door”) are important for disambiguating fine-grained parts that confuse simple part names, and MPNet’s handling of long-form text outperforms CLIP/SigLIP. Despite using  $10\times$  fewer points than PartField, we achieve superior fine-part localization (e.g., a scissors screw), confirming that semantic part-level representations are more data-efficient than dense per-point features. Key limitations are: noisy real-world scans challenge our manifold assumptions, Mahalanobis confidence degrades under distribution shift, and open-vocabulary generalization is limited to categories similar to the training data. Future work should extend this to articulated objects and integrate part-level alignments into foundation 3D models for manipulation and generation. By bridging dense geometry and structured language, ALIGN-Parts enables scalable, semantically rich 3D asset creation.



## References

- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- Minghao Chen, Roman Shapovalov, Iro Laina, Tom Monnier, Jianyuan Wang, David Novotny, and Andrea Vedaldi. Partgen: Part-level 3d generation and reconstruction with multi-view diffusion models. *arXiv preprint arXiv:2412.18608*, 2024.
- Minghao Chen, Jianyuan Wang, Roman Shapovalov, Tom Monnier, Hyunyoung Jung, Dilin Wang, Rakesh Ranjan, Iro Laina, and Andrea Vedaldi. Autopartgen: Auto-generative 3d part generation and discovery. *arXiv preprint arXiv:2507.13346*, 2025.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Jonathan H Connell and Michael Brady. Generating and generalizing models of visual objects. *Artificial Intelligence*, 31(2):159–183, 1987.
- David F. Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016.
- Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J Mitra. Diffusion 3d features (diff3f): Decorating untextured shapes with distilled semantic features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4494–4504, 2024.
- Haoan Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023.
- Aleksey Golovinskiy and Thomas Funkhouser. Randomized cuts for 3d mesh analysis. *ACM Trans. Graph.*, 27(5), 2008.
- Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: A network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4):90:1–90:12, 2019.
- R. Kenny Jones, Theresa Barton, Xianghao Xu, Kai Wang, Ellen Jiang, Paul Guerrero, Niloy Mitra, and Daniel Ritchie. Shapeassembly: Learning to generate programs for 3d shape structure synthesis. *ACM Transactions on Graphics (TOG), Siggraph Asia 2020*, 39(6):Article 234, 2020.
- Evangelos Kalogerakis, Aaron Hertzmann, and Karan Singh. Learning 3D Mesh Segmentation and Labeling. *ACM Transactions on Graphics*, 29(3), 2010.
- Evangelos Kalogerakis, Melinos Averkiou, Subhransu Maji, and Siddhartha Chaudhuri. 3D Shape Segmentation with Projective Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6630–6639, Los Alamitos, CA, USA, 2017. IEEE Computer Society.
- Hyunjin Kim and Minhyuk Sung. Partstad: 2d-to-3d part segmentation task adaptation. In *ECCV*, 2024.
- 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. *arXiv:2304.02643*, 2023.
- Yuchen Lin, Chenguo Lin, Panwang Pan, Honglei Yan, Yiqiang Feng, Yadong Mu, and Katerina Fragkiadaki. Partrafter: Structured 3d mesh generation via compositional latent diffusion transformers, 2025.
- Minghua Liu, Yinhao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, and Hao Su. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21736–21746, 2023.
- 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. *arXiv preprint arXiv:2504.11451*, 2025.
- Changfeng Ma, Yang Li, Xinhao Yan, Jiachen Xu, Yunhan Yang, Chunshi Wang, Zibo Zhao, Yanwen Guo, Zhuo Chen, and Chunchao Guo. P3-sam: Native 3d part segmentation, 2025.
- Ziqi Ma, Yisong Yue, and Georgia Gkioxari. Find any part in 3d. *arXiv preprint arXiv:2411.13550*, 2024.
- 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.
- Lior Shapira, Ariel Shamir, and Daniel Cohen-Or. Consistent mesh partitioning and skeletonisation using the shape diameter function. *Vis. Comput.*, 24(4):249–259, 2008.
- Habib Slim, Xiang Li, Yuchen Li, Mahmoud Ahmed, Mohamed Ayman, Ujjwal Upadhyay, Ahmed Abdelreheem, Arpit Prajapati, Suhail Pothigara, Peter Wonka, and Mohamed Elhoseiny. 3DComPaT++: An improved large-scale 3D vision dataset for compositional recognition. *arXiv preprint*, 2023.
- 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.
- Karl Stelzner, Kristian Kersting, and Adam R. Kosiorek. Decomposing 3d scenes into objects via unsupervised volume segmentation. *ArXiv*, abs/2104.01148, 2021.
- Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- George Tang, William Zhao, Logan Ford, David Benhaim, and Paul Zhang. Segment any mesh. *arXiv preprint arXiv:2408.13679*, 2024.
- Jiaxiang Tang, Ruijie Lu, Zhaoshuo Li, Zekun Hao, Xuan Li, Fangyin Wei, Shuran Song, Gang Zeng, Ming-Yu Liu, and Tsung-Yi Lin. Efficient part-level 3d object generation via dual volume packing. *arXiv preprint arXiv:2506.09980*, 2025.
- Shubham Tulsiani, Hao Su, Leonidas J. Guibas, Alexei A. Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Ardian Umam, Cheng-Kun Yang, Min-Hung Chen, Jen-Hui Chuang, and Yen-Yu Lin. Partdistill: 3d shape part segmentation by vision-language model distillation. In *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2024.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- 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.
- Yunhan Yang, Yuan-Chen Guo, Yukun Huang, Zi-Xin Zou, Zhipeng Yu, Yangguang Li, Yan-Pei Cao, and Xihui Liu. Holopart: Generative 3d part amodal segmentation. *arXiv preprint arXiv:2504.07943*, 2025.
- Yibo Zhang, Li Zhang, Rui Ma, and Nan Cao. Texverse: A universe of 3d objects with high-resolution textures. *arXiv preprint arXiv:2508.10868*, 2025.
- Yuchen Zhou, Jiayuan Gu, Xuanlin Li, Minghua Liu, Yunhao Fang, and Hao Su. Partslip++: Enhancing low-shot 3d part segmentation via multi-view instance segmentation and maximum likelihood estimation. *arXiv preprint arXiv:2312.03015*, 2023.

## A Experiments and Analysis

Given the challenges inherent in semantic 3D part segmentation, we find that no current published work is directly comparable to our method. To enable rigorous evaluation, we introduce our own strong baseline detailed in Section A.1. While we do include comparisons against class-agnostic 3D part segmentation methods in this manuscript, it is important to note that these do not constitute an entirely fair benchmark for our approach. Most prior methods have been trained using proprietary, closed-source Objaverse-scale datasets, with specific data details and part annotations rarely disclosed publicly.

In contrast, our experiments are conducted on fully open, publicly available datasets, and our methodology itself improves upon these resources, making our results more easily reproducible and comparable for future researchers. Furthermore, a key emphasis of our approach is efficiency: we process only 10,000 input points per shape, in stark contrast to the 100,000 points typically used by class-agnostic segmentation baselines. This restriction stems from the academic compute limitations we faced, while prior works often benefit from corporate-scale GPU resources.

Despite these constraints, our method achieves competitive or superior performance relative to existing baselines. It is reasonable to expect that - if provided with similar data volume and computational resources - ALIGN-Parts would further extend its advantage on standard metrics and benchmarks. Our design choices thus not only democratize research in 3D part segmentation but also highlight the promise of reproducibility, accessibility, and efficiency for large-scale semantic understanding in open 3D datasets.

### A.1 PartField+MPNet baseline

Given that our task of semantic part segmentation (in contrast to the relatively easier and more prominent class agnostic part segmentation, we create our own baseline - PartField + MPNet, which assigns labels to parts obtained by KMeans clustering on per-point features. We experimented with two variants of this model, in terms of input features: PartField and PartField + DINOv2, and found that the latter usually yields much better performance. So, without loss of generality, our baseline PartField + MPNet refers to the model where we have per-part PartField + DINOv2 features fused through cross-attention. Specifically, we employ cross-attention fusion to combine per-part geometric (448-D) and appearance (768-D) features, projecting them through 512-D hidden layers into a shared 256-D latent space. The architecture consists of a dense feature fuser (2.8M parameters) with 4 attention heads operating at 512-D, followed by dedicated MLP projectors for local part features (0.39M), semantic text embeddings (0.52M), and global shape descriptors (0.75M), totaling approximately 5.1M parameters. Training optimizes three objectives: symmetric InfoNCE loss for local part-text alignment, a global-level contrastive loss between shape and class embeddings, and a cross-entropy clustering loss that predicts part counts with equal weighting ( $\lambda=1.0$ ) across all terms. The model is trained for 100 epochs using AdamW with learning rate  $3e-4$ , weight decay  $1e-5$ , and cosine annealing schedule ( $\eta_{\min}=5e-6$ ) with batch size 64. The part count prediction head (0.63M parameters) uses a two-layer MLP with GELU activation to classify the number of semantic parts from fused global features. All projectors and

attention mechanisms utilize dropout regularization ( $p = 0.1$ ) to prevent overfitting during training. During inference, PartField + MPNet first predicts object category by comparing the projected global feature against all class embeddings, then performs soft k-means clustering ( $k$  from the part count head) on fused point-level features with Hungarian matching to assign semantic labels by computing cosine similarity between projected cluster centroids and MPNet embeddings of candidate part names.

### A.2 Part-Retrieval Comparison with Find3D

Beyond semantic segmentation, ALIGN-Parts also supports text-driven part retrieval—the task of localizing and retrieving point cloud regions corresponding to natural language part queries. This capability, introduced by Find3D, enables flexible, open-vocabulary part discovery directly from unstructured text descriptions. Our approach performs retrieval by constraining the candidate label vocabulary to only those parts known to be present in the target object class, rather than the full semantic vocabulary. Additionally, we set the number of active partlets to match the ground-truth part count for the object, which serves as an oracle constraint. While this restriction reduces the search space and assignment ambiguity, allowing the model to match predicted part slots to a small, object-specific set of valid labels rather than choosing from dozens of candidates, it also enables fairer and more interpretable comparisons.

This constrained retrieval setup typically yields higher segmentation accuracy by minimizing false positive label assignments and focusing the model’s attention on semantically coherent parts. The key advantage of our approach lies in the compositional three-level hierarchy: point cloud  $\rightarrow$  partlet  $\rightarrow$  part label. This formulation naturally encourages the discovery of connected point groups with consistent semantic meaning, whereas alternatives may suffer from fragmentation or over-segmentation.

We present qualitative comparisons with Find3D on two representative 3D objects from the *airplane* and *motorbike* object classes in the Objaverse-General benchmark (part of our closed-vocabulary evaluation set), released by the Find3D authors. As shown in Figure 13, ALIGN-Parts consistently retrieves more spatially coherent and semantically meaningful part groups, demonstrating the effectiveness of our partlet-based design for part localization and retrieval tasks.

**Figure 13: Part Retrieval Comparison with Find3D.** We demonstrate text-driven part retrieval on two representative objects (airplane and motorbike) from Objaverse-General. Given natural language part queries (e.g., “body”, “wing”, “gas tank”, “wheel”), ALIGN-Parts identifies and retrieves spatially coherent point groups corresponding to each part. Compared to Find3D (left), our method produces more semantically and spatially consistent part retrievals by leveraging the hierarchical point  $\rightarrow$  partlet  $\rightarrow$  part label decomposition. This design encourages the discovery of well-connected, semantically meaningful regions rather than fragmented point clusters. Ground truth part segmentations (right) show the target labels. ALIGN-Parts achieves results that closely align with ground truth, validating the effectiveness of our partlet-based formulation for open-vocabulary part localization and retrieval.

