Vision-Language Model Enhanced Semantic Part Labeling

Vivian Lu
Brown University
vivian_f_lu@brown.edu

Abstract

SEGRUE (SEGmenting 3D Shapes from RUles and Examples) is a system designed to assign semantic part labels to the regions of pre-segmented 3D shapes, a task with broad applications in areas such as CAD modeling and robot-object interaction. To perform labeling, SEGRUE uses a set of label functions, each dedicated to a specific label in a predefined grammar (e.g., `chair_arm`, `back_surface`). These functions evaluate the geometric and spatial characteristics of each region and choose whether or not to "fire" (assign a label). The weak labels produced by these functions are used to train a neural network that learns to predict semantic labels for new, unseen shape regions.

My contributions to this project explore methodologies for designing and constructing effective label functions by incorporating a vision-language model (VLM) into the pipeline. While the existing system relies on a shape analysis API that focuses on low-level geometric primitives, my approach uses natural language prompts paired with rendered images of object components to query the VLM. This enables the assessment of not only geometric properties but also relational positioning and functional characteristics that are challenging to capture with traditional geometric methods alone.

1 Background

This work relies on PartNet [2], for a hierarchical dataset of 3D object parts, and builds on NGSP [1], which introduces a grammar-based approach for part labeling. We assume the 3D objects provided as input to the system have been segmented into distinct, unlabeled regions. However, assigning consistent and accurate semantic part labels remains a challenging task due to wide variations in geometry and spatial layout (see Figure 1). To address this, SEGRUE relies on interpretable label functions constructed over an API that supports reasoning about low-level geometric properties and spatial relationships between regions.

Figure 1. Examples of 3D chair models from PartNet.

The `segrue_api` defines a set of functions that operate on regions, each supporting either geometric reasoning based on the region's mesh structure or relational reasoning based on its spatial context. For example, a geometric function like `get_aspect_ratio` determines the relative scale of each dimension:

```
def get_aspect_ratio(region: Region) -> List
    [float]:
    """
    Returns the normalized aspect ratio
    (width,height,depth) of the region,
    normalized so the largest value is 1.0.
    """
```

Conversely, relational functions like `is_supporting` examine how regions interact with each other:

```
def is_supporting(query_region: Region,
    key_region: Region) -> bool:
    """
    Returns whether query_region provides
    physical support for key_region.
    """
```

Label functions combine these API functions to determine when a region matches a specific semantic label from the grammar. For example, this simplified label function aims to fire when the provided region is a seat surface:

```
def seat_surface_label_function(region,
    label="chair/chair_seat/seat_surface"):
    centered =
        get_distance_to_center(region) < 0.1
    supports_back =
        is_supporting(region, back_region)
    if centered and supports_back:
        return label
    else:
        return None
```

In the process of developing effective label functions, we first establish a baseline by handwriting label functions using the functions provided in the `segrue_api`. Once we achieve reasonable performance with these handwritten functions, the goal is to leverage large language models to automatically generate label functions. The ultimate goal is for these LLM-generated label functions to outperform their handwritten counterparts and produce more accurate weak labels.

However, the geometric and relational primitives provided by the `segrue_api` have inherent limitations in capturing certain semantic characteristics that humans can intuitively recognize. This includes visual features that are difficult to quantify geometrically and functional properties that relate to how humans may interact with an object. My contributions explore how vision-language models can augment this existing framework, enabling label functions to leverage visual reasoning capabilities beyond low-level geometric primitives.

2 Method

In this section, I detail my approach to enhancing semantic part labeling through the integration of a vision language model into the system. I begin by exploring how VLMs can assess functional properties of object parts—moving beyond purely geometric reasoning. This culminates in two complementary approaches for chair part labeling: a function-based global question set and comprehensive region-specific question sets designed to establish the upper bound for label function performance when integrating a vision-language model.

I explored how designing a new API centered around queries to a VLM might enable more capabilities in semantic labeling. In particular, if our label functions could consist of a set of queries to a VLM rather than relying on an API that purely analyzed low-level geometric properties, we could potentially produce higher-quality weak labels and correctly label regions we were missing previously.

2.1 Exploring Functional Properties via VLM

Building on the initial API extensions, I next investigated a more fundamental shift in approach: developing an entirely new set of VLM-based functions focused on functional properties rather than geometric ones. While the previous extensions used the VLM as a computational proxy for geometric analysis, this approach explores whether VLMs could directly assess higher-level functional characteristics that are difficult to capture through conventional geometric reasoning, asking questions like "Is this region meant to block light?" or "Is region designed to be gripped for handling the object?".

This approach was tested across a diverse set of object categories from PartNet, including tables, chairs, lamps, beds, knives, scissors, mugs, vases, bottles, bowls, displays, laptops, earphones, and hats. This broad selection allowed for testing how well the prompts worked across different object types and whether the VLM could recognize similar functional properties across different object types (e.g., identifying that both a chair arm and bed headboard are "designed to be leaned against").

The goal of this exploration was to determine whether natural language prompting about functional properties could provide a viable alternative or complement to traditional geometric analysis. Rather than attempting to replicate geometric measurements, this approach leverages and measures the VLM's ability to directly assess how different parts of objects are meant to be used, which may be intuitive to humans but challenging to express through geometric primitives.

2.2 Function-Based Question Set for Enhanced Chair Labeling

After confirming the VLM's ability to reason about functional properties across diverse object categories, I narrowed my focus to improving performance specifically on chair part labeling. This decision allowed for direct comparison with the label functions designed by the original `segrue_api` to determine whether an approach centered around VLM queries could produce more effective label functions than those based on geometric analysis.

I designed a "global question set" of functional questions, each designed to target specific regions. While these questions were intentionally targeted, they were formulated to be broad enough to apply to multiple regions with similar functional roles. For example, the question "Is this part designed to maintain the overall stability of the object, preventing it from tipping over?" was designed to identify multiple components that contribute to stability, including `bar_stretcher`, `foot`, `leg`, and `runner` regions. Other questions addressed different functional aspects, such as "Is this part designed to provide upper body support if a person were to sit on the object?".

Using these functional questions, I constructed label functions for each chair component by testing multiple functional properties. For example, a label function for the leg might look like:

```
def leg_label_function(region,
    label="chair/chair_base/leg"):
    ...
    if provides_structural_support and
        maintains_vertical_support and
        keeps_object_stable:
        return label
    else:
        return None
```

This approach represents a fundamental shift in how label functions are defined. Rather than using geometric properties like dimensions, orientation, or position, these functions identify parts based on their functional role within the object as a whole. The combinatorial logic allows for precise differentiation between parts that share some functional properties but differ in others. For instance, while both `back_surface` and `seat_surface` are designed to provide comfort, they support different parts of the body.

By formulating label functions in terms of functional properties, this method leverages the VLM's understanding of

object-human interaction rather than relying on explicit geometric measurements, potentially making the approach more robust to variations in region shapes.

2.3 Comprehensive Question Sets for Upper-Bound VLM Performance

Building upon the function-based question set discussed in the previous section, I next tested a more specialized strategy aimed at maximizing the potential performance of VLM-based label functions. This approach involved creating targeted question sets customized for each specific chair region, with the goal of estimating an upper bound on what could be achieved with VLM-based reasoning.

Unlike the previous global question set that focused exclusively on functional properties, these targeted questions were designed to leverage three complementary categories:

1. **Geometric Properties**: questions about shape, form, orientation, and physical characteristics
2. **Functional Properties**: questions about purpose and intended use
3. **Relational Properties**: questions about connections and positions relative to other chair components

The inclusion of relational properties represents a particularly important advantage of VLM-based approaches. When asking relational questions like "Is this part surrounding the back surface?" the VLM can directly assess the visual relationship without depending on prior labeling. This is possible because the VLM should recognize what constitute a back surface based on its appearance, position, and context within the chair, and then determine whether the highlighted region surrounds it. This contrasts with the original `segrue_api` approach, where relational functions like `is_surrounding(back_frame, back_surface)` would depend entirely on the accuracy of previously identified back surface regions. In that approach, the back surface regions would first need to be identified by a separate back surface label function, which itself might contain errors or inaccuracies.

For each chair region, I wrote a unique set of 3-4 questions designed to create a distinct set of responses that would uniquely identify that region. This combination of questions was intended to formulate a label function to target one region specifically, and in the ideal scenario firing only for the intended region and no others. For example, for the `chair_arm` region, the questions are:

1. Is this part connected to the back surface?
2. Is this part positioned above and to the side of the seat?
3. Is there another similarly shaped part mirrored on the opposite side of the chair?

For the `leg` region:

1. Is this part connected directly to the underside of the seat?
2. Is this part in direct contact with the ground?
3. Does this part extend from the sat to the ground without interruption?
4. Is this part responsible for supporting the vertical weight of the seat and user?

This approach represents a significant refinement over the previous function-based question set by tailoring each question to precisely distinguish one region from all others. While the previous approach tested how functional properties could identify chair components, this targeted approach is meant to explore the upper bound of what could be achieved when combining functional properties with geometric and relational characteristics in a region-specific manner.

3 Results

This section presents the evaluation of our VLM-enhanced approaches compared to the baseline geometric method. We define three scenarios for comparison:

- **Scenario A:** The baseline approach using traditional geometric-based label functions in the `segrue_api`.
- **Scenario B:** Our function-based global question set approach that leverages VLM for functional properties.
- **Scenario C:** Our comprehensive region-specific question sets approach that combines geometric, functional, and relational properties.

The performance of label functions in each scenario is evaluated by running the label functions on the first 100 chairs from the training set. Every label function is run on every region, and each time a label function fires the fired label is compared to the ground truth annotation from PartNet. Based on this data, precision, recall, F1, and accuracy are calculated for each label function.

3.1 Function-Based Global Question Set for Enhanced Chair Labeling

This section evaluates the performance of our function-based global question set approach, which uses VLM queries focused on functional properties as described in Section 2.2. We compare the performance of label functions in Scenarios A and B as shown in Table 1 and Table 2.

Region	Precision A	Precision B	Recall A	Recall B
chair_arm	1.00	0.51	0.14	0.26
back_frame	0.38	0.45	0.04	0.29
back_surface	0.12	0.78	0.76	0.32
bar_stretcher	0.41	0.39	0.12	0.12
leg	1.00	0.38	0.45	0.12
runner	0.00	0.51	0.00	0.37

Table 1. Precision and recall for each region under scenario A and scenario B, comparing label function performance.

Region	F1 A	F1 B	Accuracy A	Accuracy B
chair_arm	0.25	0.35	0.89	0.87
back_frame	0.07	0.35	0.92	0.91
back_surface	0.21	0.45	0.46	0.88
bar_stretcher	0.19	0.19	0.87	0.93
leg	0.62	0.18	0.81	0.77
runner	0.00	0.42	0.88	0.94

Table 2. F1 score and accuracy for each region under scenario A and scenario B, comparing label function performance.

The VLM-based approach demonstrates significant improvements in F1 scores across nearly all chair regions. Particularly notable are the increases for `back_frame` (from 0.07 to 0.35) and `runner` (from 0.00 to 0.42), with the latter region being completely undetectable using the geometric baseline approach. Similarly, `back_surface` shows a substantial improvement in F1 score (from 0.21 to 0.45) and accuracy (from 0.46 to 0.88), demonstrating the effectiveness of functional questions in identifying this region.

The only exception is the `leg` region, where the F1 score decreases from 0.62 to 0.18. This anomaly likely stems from the specific implementation of the leg label function rather than an inherent limitation of the VLM approach, as the VLM consistently improves performance for all other regions. Furthermore, this isolated case stands in contrast to the overall pattern of improved accuracy metrics across the board, suggesting that with refinement, the leg label function could achieve comparable or superior performance to the geometric approach.

3.2 Comprehensive Region-Specific Question Sets

This section evaluates the performance of our region-specific question sets approach described in Section 2.3, which uses tailored combinations of geometric, functional, and relational properties for each chair component. The tables below compare the baseline geometric approach (Scenario A) with our comprehensive region-specific question sets (Scenario C).

Region	Precision A	Precision C	Recall A	Recall C
chair_arm	1.00	0.55	0.14	0.90
back_frame	0.38	0.38	0.04	0.09
back_surface	0.12	1.00	0.76	0.03
bar_stretcher	0.41	0.64	0.12	0.50
leg	1.00	0.93	0.45	0.47
rocker	0.00	0.52	0.00	0.55
runner	0.00	0.92	0.00	0.10
seat_frame	0.00	0.13	0.00	0.52
seat_surface	0.80	0.94	0.15	0.15

Table 3. Precision and recall for each region under scenario A and scenario C, comparing label function performance.

Region	F1 A	F1 C	Accuracy A	Accuracy C
chair_arm	0.25	0.69	0.89	0.92
back_frame	0.07	0.15	0.92	0.93
back_surface	0.21	0.06	0.46	0.91
bar_stretcher	0.19	0.56	0.87	0.91
leg	0.62	0.62	0.81	0.80
rocker	0.00	0.53	0.98	0.98
runner	0.00	0.18	0.88	0.90
seat_frame	0.00	0.21	0.95	0.80
seat_surface	0.26	0.26	0.88	0.92

Table 4. F1 score and accuracy comparison between Scenario A and Scenario C, comparing label function performance

For several regions, the approach yields dramatic improvements in F1 scores: `chair_arm` shows a 176% increase (from 0.25 to 0.69), `bar_stretcher` improves by 195% (from 0.19 to 0.56), and `rocker` goes from completely undetectable (F1 = 0.00) to a robust F1 score of 0.53. Additionally, the VLM approach is capable of identifying regions that were almost undetectable using the label functions based on the original `segrue_api` such as `rocker`, `runner`, and `seat_frame`.

For `back_surface`, our approach in scenario C achieves perfect precision (1.00 vs. 0.12) but at the cost of severely reduced recall (0.03 vs. 0.76), suggesting our questions may be too restrictive. Notably, despite the lower F1 score, the accuracy for this region improves substantially from 0.46 to 0.91, indicating the approach makes fewer errors overall across the dataset. This suggests that while our method identifies fewer back surfaces, the ones it does identify are correct with high confidence, and it correctly excludes non-back-surface regions.

Overall, the accuracy metrics show improvements for seven out of nine regions, with the most substantial increases for `back_surface` (+0.45) and more modest gains for regions like `chair_arm` (+0.03) and `bar_stretcher` (+0.04).

4 Conclusion

This work demonstrates that integrating vision-language models into the semantic part labeling pipeline shows potential improvements in performance over traditional geometric methods. Both global functional questions and region-specific question sets show gains in F1 score and accuracy, particularly for parts with clear functional roles or those that were previously difficult to detect. However, the effectiveness of these approaches can vary across regions, and some declines in performance (such as reduced recall for certain regions) highlight the need for further refinement and experimentation with various prompting strategies. Moving forward, two key extensions can significantly enhance this pipeline:

1. **LLM-generated VLM prompts:** The current results reported are based on manually authored question sets and label functions for each region. The ultimate goal

is to be able to automate this process using LLMs and have the LLM produce label functions that can match the performance of handwritten label functions.

2. **Improved region renderings:** The current reliance on a fixed rendering angle provided by PartNet limits the performance of the prompts supplied to the VLM. PartNet renders all region images in an upper-front left view (see Figure 2) with the targeted region highlighted in red. However, this limits visibility of certain parts, especially small or occluded regions such as feet, chair arms, and bar stretchers. Rendering regions from multiple viewpoints or dynamically selecting angles that maximize visibility of the target region would likely improve VLM performance. For instance, symmetry-related questions about parts like chair arms are much harder to answer without a head-on perspective.

Figure 2. Examples of foot, rocker, and bar stretcher renderings from PartNet. The fixed viewpoint limits visibility of small or occluded parts.

5 Acknowledgements

The report is part of a larger research project explored with Jean Yoo, Kenny Jones, and Daniel Ritchie.

References

[1] R. Kenny Jones, Aalia Habib, Rana Hanocka, and Daniel Ritchie. 2022. The Neurally-Guided Shape Parser: Grammar-based Labeling of 3D Shape Regions with Approximate Inference. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).

[2] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. 2019. 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).*