CoSMo3D: Open-World Promptable 3D Semantic Segmentation through LLM-Guided Canonical Spatial Modeling

Supplementary Material

1. Outline

In this supplementary material, we provide additional details and analyses that complement the main paper, including: (1) further details for the cross-category canonicalization pipeline (Section 2); (2) extended analysis of the ablation studies (Section 3); and (3) additional qualitative comparisons (Section 4).

2. More Canonicalization Details

We achieve cross-category object canonicalization in two steps: first, using GPT to cluster a large number of categories, and second, performing cross-category alignment based on the resulting clusters.

GPT-Assisted Clustering: This step realizes category clustering by invoking GPT with two prompts, as shown in Table S1. First, Prompt 1 is used to obtain the names of a cluster list in the open world. Subsequently, Prompt 2 is applied to determine the cluster affiliation of each of the 200 categories in the dataset—i.e., which cluster each category belongs to. The cluster affiliation judgment strictly adheres to the principles of core semantic components and consistent functional alignment to avoid ambiguity and inter-cluster overlap. The information of the finally extracted clusters is presented in Table S2.

Cross-Category Semantic and Functional Alignment. Building on the category clusters obtained in the previous step, we perform cross-category alignment via a lightweight manual procedure. For each cluster, we first select one object from a chosen category as the alignment template, which serves as the reference for aligning objects from the remaining categories in the same cluster. Using MeshLab, we then adjust the poses (i.e., rotation angles) of representative objects from the remaining categories so that their orientations are consistent with the template in terms of both semantic direction and functional usage.

After completing intra-cluster cross-category alignment, we further perform cross-cluster alignment using the same template selection and pose adjustment strategy, ultimately constructing a unified canonical space across all categories. Since the underlying dataset has already undergone intra-category alignment (i.e., objects within each category share a canonical pose), we only need to align one representative object per category to determine the pose offset applied to the entire category. In practice, aligning 200 categories took approximately 4 hours, which we find sufficient for practical applications.

Table S1. Customized Prompts for Cluster Extraction and Category Affiliation Determination

Two Core Customized Prompts	
Prompt 1: Major Cluster Extraction	**Prompt 2: Category Affiliation Determination**
Taking "functional similarity of objects" and "semantic relevance of usage" as the core criteria, cluster object categories in the open world to generate 15-20 semantic clusters. Each cluster should be named in the structure of "function + component/port". Output format requirement: Arrange in two columns, mark with serial numbers, and present cluster names in italics. Examples: 1. *sound output port* 2. *flame output port*	Given the existing semantic cluster listCluster name list, please judge the affiliation of the Category name based on the principles of "core semantic component matching" and "functional usage alignment". Output format: Category name + corresponding cluster number and name, ensuring no ambiguity and no overlap.

Table S2. GPT-based semantic clustering results. Each cluster groups object categories with similar functionality and usage semantics.

Semantic Cluster List from GPT	
1. *sound output port*	2. *flame output port*
3. *fluid air port*	4. *light output port*
5. *weapon attack part*	6. *manual tool part*
7. *electronic operate part*	8. *instrument play part*
9. *general container part*	10. *plant container part*
11. *bath container part*	12. *wearable fit part*
13. *rotate axis part*	14. *arch fixed part*
15. *measure display part*	16. *special function part*
17. *furniture support part*	18. *transport carry part*
19. *heat electric part*	

3. More Ablation Analysis

To intuitively illustrate the contribution of each proposed module, we present qualitative visualizations corresponding to the ablation variants discussed in the main paper. For quick reference, the configuration of each ablation variant is summarized in Table S3.

Table S3. Configuration of ablation variants (corresponding to Table 2 in the main paper). Each column indicates which modules are enabled in each variant.

Module	A	B	C	D	Full Model
Hard-Negative Sampling		✓	✓	✓	✓
Canonical Map Anchoring			✓	✓	✓
Cross-Cat. Canon. (Data)				✓	✓
Canonical Box Calibration					✓

* All settings are trained on intra-category canonicalized shapes.
* 'Cross-Cat. Canon (Data)' denotes processing the dataset via our cross-category canonicalization pipeline.

Hard-Negative Sampling. As shown in Fig. S1, we visualize the effect of enabling the Hard-Negative Sampling strategy (variant B vs. variant A). In the

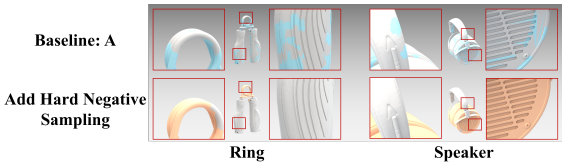

Figure S1. **Ablation Visualization of Hard Negative Sampling.** Comparison between the baseline without Hard-Negative Sampling (variant A) and the model with Hard-Negative Sampling enabled (variant B). The baseline suffers from noisy segmentation and imprecise part boundaries. Adding the proposed sampling strategy improves edge clarity and reduces feature noise, especially around thin or complex structures (e.g., ring boundaries, speaker grills).

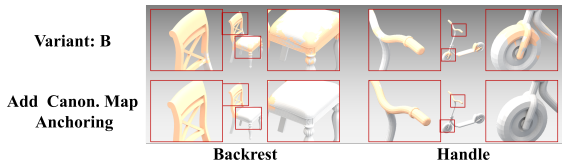

Figure S2. **Ablation Visualization of Canonical Map Anchoring.** Top: variant with contrastive loss only (variant B). Bottom: variant with canonical map anchoring loss added (variant C). Geometrically similar parts, such as a chair's backrest and seat or a bicycle's handle and wheel, are often misclassified without anchoring. Canonical map anchoring loss provides spatial priors in canonical space, enabling correct semantic separation.

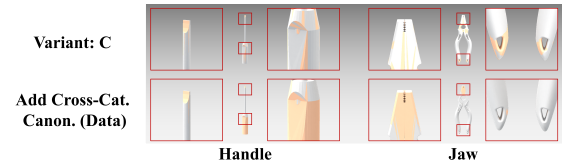

Figure S3. **Ablation Visualization of Cross-Category Canonicalization.** Top: model trained with intra-category canonical data only (variant C). Bottom: model trained with unified cross-category canonical data (variant D). Without cross-category alignment, same-semantic parts (e.g., "handle", "jaw") are mapped inconsistently across object types, resulting in segmentation conflicts. Cross-category canonicalization resolves these issues by aligning shared semantics spatially.

baseline setting (variant A), contrastive learning is applied between the average features of parts and their associated text embeddings, without explicitly constraining the feature variance within each part. This leads to noisy predictions and unclear boundaries. When hard negative sampling is introduced (variant B), the model learns to emphasize discriminative regions, especially near part boundaries. As a result, the segmentation becomes more consistent, with reduced noise and sharper edges. Improvements are particularly noticeable on geometrically complex structures such as rings and speaker grills, as shown in Fig. S1.

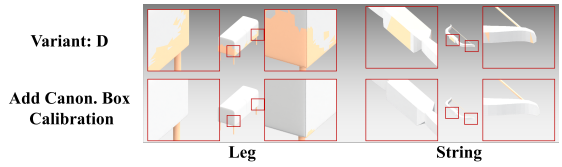

Figure S4. **Ablation Visualization of Canonical Box Calibration.** Top: model without Canonical Box Calibration (variant D). Bottom: with Canonical Box Calibration (Full Model). This constraint enhances spatial reasoning in canonical space, produces sharper part boundaries (e.g., leg, string).

Canonical Map Anchoring. As Fig. S2 shows, when supervision is based solely on the contrastive alignment loss, which primarily relies on local geometry and text semantics (variant B), the model struggles to differentiate parts with similar shapes but distinct semantic roles. For example, the backrest and seat of a chair both resemble flat cuboids, and the handle and wheel of a bicycle are both cylindrical in geometry. This often leads to incorrect segmentation, as the model confuses geometrically similar but semantically different parts. By introducing the canonical map anchoring loss (variant C), the model learns to associate parts with their relative positions in canonical space, providing a global spatial prior. This results in more reliable segmentation: despite ambiguity in local geometry, the consistent semantic layout in canonical space guides the model toward correct part identification.

Cross-Category Canonicalization. As visualized in Fig. S3, we compare variant C (trained using intra-category canonicalized shapes only) with variant D (which additionally incorporates cross-category canonical alignment). While the canonical map anchoring loss introduced in variant C helps separate semantically distinct parts with similar geometry, it still suffers from inconsistent canonical-space placement across categories—since intra-category alignment alone cannot ensure that identical semantics are consistently located across different object types. For example, parts labeled as "handle" or "jaw" may appear in misaligned regions, leading to segmentation conflicts. By applying cross-category canonicalization to training data (variant D), semantically identical parts are mapped to more consistent locations in canonical space. This improves overall segmentation consistency and resolves ambiguities for shared parts across categories.

Canonical Box Calibration. As shown in Fig. S4, the addition of the canonical box calibration loss (Full Model vs. variant D) further enhances the model's perception of canonical space. By explicitly supervising the spatial extent of each part, this constraint sharpens part boundaries and reduces ambiguity near

region edges. For example, the "leg" and "string" regions become more clearly delineated when this loss is applied. As a result, the full model yields more precise and semantically coherent segmentation, even in fine-grained or spatially adjacent parts.

4. More qualitative experiments

To further demonstrate the effectiveness of our method, we present additional qualitative comparisons against the state-of-the-art approach Find3D [1], as shown in Fig. S5. We evaluate both methods across a wide range of shapes and text prompts. Find3D relies primarily on local geometric-textual alignment, where part-level feature pooling is used. However, this often leads to noisy predictions and confusion between geometrically similar parts. In contrast, our method enhances canonical space perception through the proposed modules, introducing a global spatial prior, enabling more consistent semantic understanding. This effectively reduces ambiguity in cases where local geometry alone is insufficient to resolve part semantics (e.g., distinguishing "handle" from "cord" or "plant" stems). Moreover, our design avoids the significant noise caused by relying solely contrastive loss applied to average-pooled part features.

References

[1] Ziqi Ma, Yisong Yue, and Georgia Gkioxari. Find any part in 3d. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7818–7827, 2025. 3

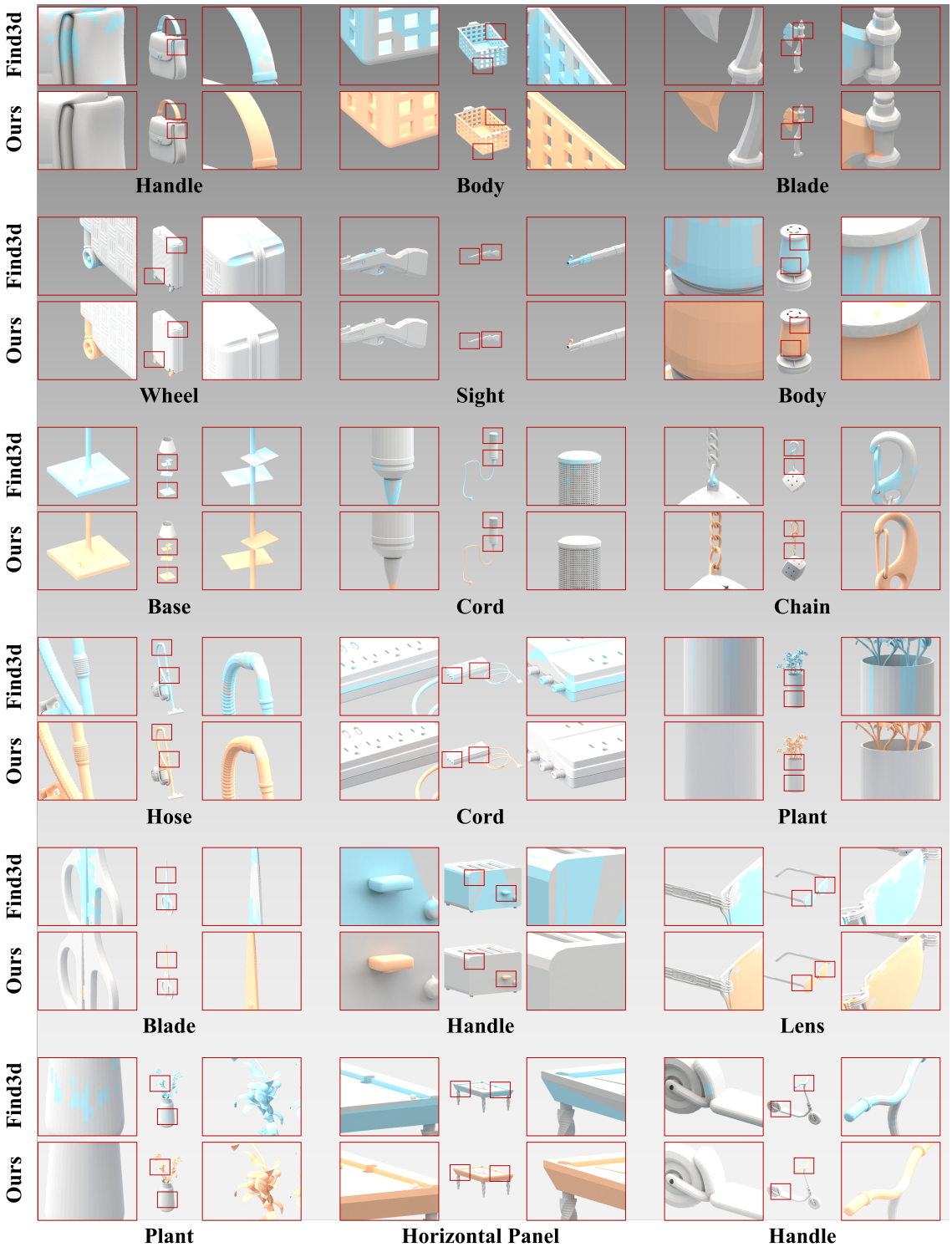

Figure S5. More Qualitative Experiments: Comparison between Our Method (in Orange) and the State-of-the-Art Method Find3D (in Blue). Our method achieves more accurate segmentation with less noise.