OSCAR: Open-Set CAD Retrieval from a Language Prompt and a Single Image

Tessa Pulli[1], Jean-Baptiste Weibel[2], Peter Hönig[1], Matthias Hirschmanner[1], Markus Vincze[1]

name@acin.tuwien.ac.at

[1]Automation and Control Institute, TU Wien, Wien, Austria
[2]BOKU University, Institute for Forest Engineering, Wien, Austria

Problem Statement

- **6D object pose estimation** is key for object manipulation in robotics
- Training-based methods fail on **unseen objects**

⇨ **Zero-shot methods**

- Do not require training but rely on a **3D object model for each instance**

Option 1:

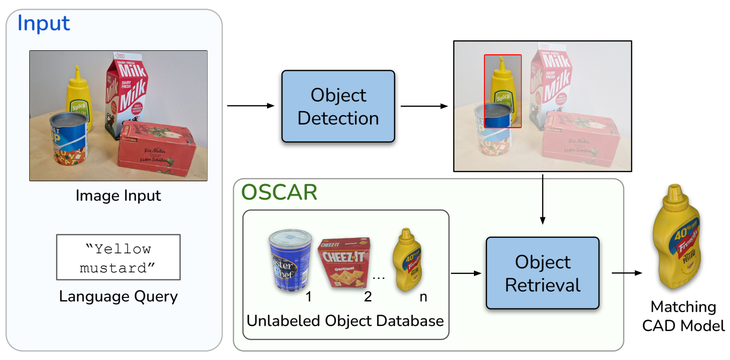
- Manage database of object models manually → Prone to human errors

Option 2:

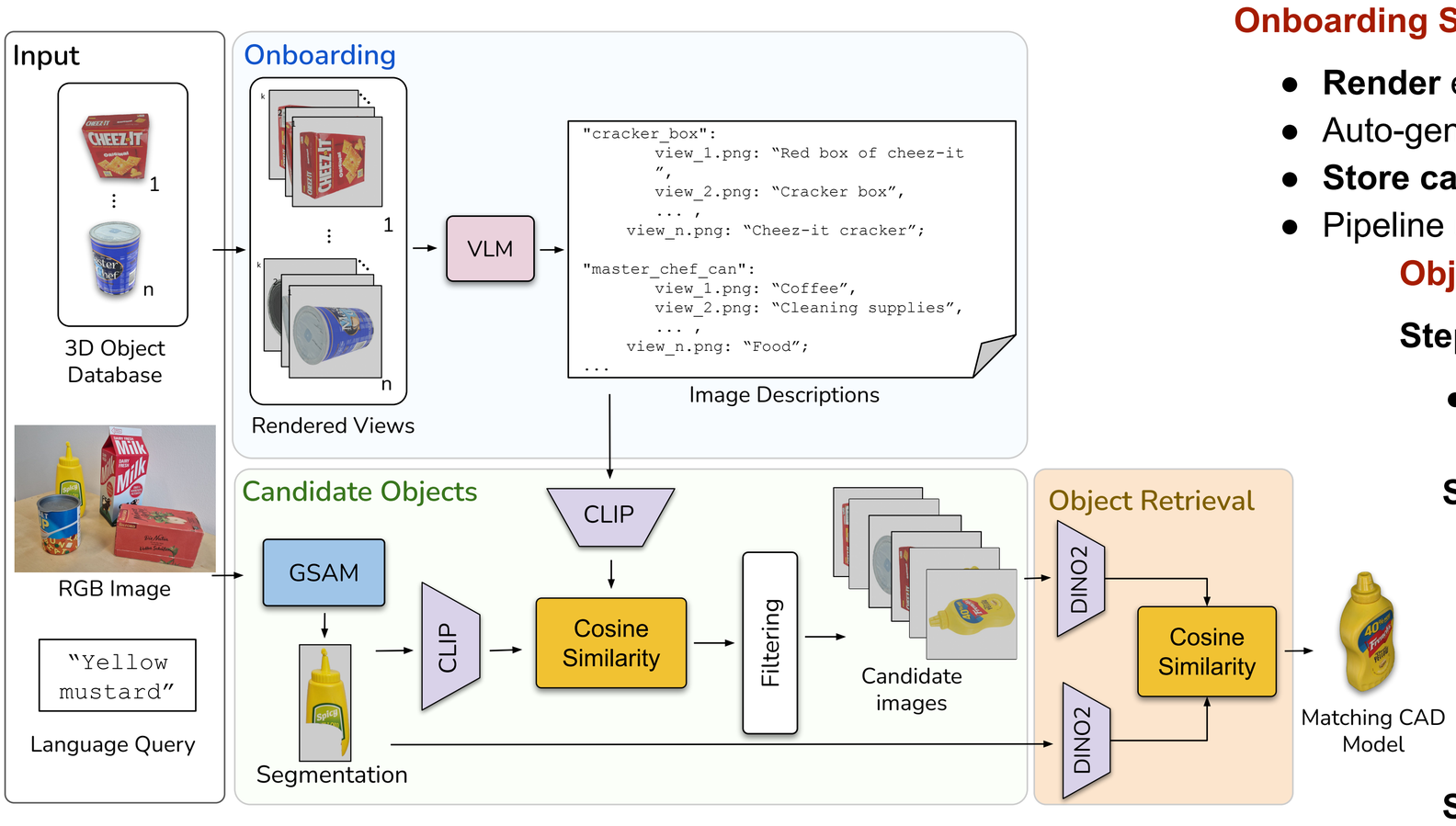
- Reconstruct objects on the fly → computational expensive, poor 3D models, time-intensive

Contribution

- OSCAR retrieves the matching CAD model from an unlabeled 3D database using an RGB image + language prompt. → **Database Management**
- Alternative approach to **zero-shot pose estimation** when the ground-truth model is not available.



Method



Overview of OSCAR: In the onboarding stage, CAD models are rendered from multiple viewpoints and automatically captioned. At inference, an object ROI is segmented from the input image using GroundedSAM. Retrieval is performed in two stages: (1) text-based filtering with CLIP, and (2) image-based matching with DINOv2, yielding the most similar CAD model for pose estimation.

Onboarding Stage:

- **Render** each CAD model from 8 viewpoints
- Auto-generate **captions with LLaVA**
- **Store captions + rendered** views for retrieval
- Pipeline updates automatically when new models are added

Object Retrieval

Step 1: ROI Extraction

- Use GroundedSAM to segment the queried object

Step 2: Candidate Filtering (Text-based)

- Encode ROI with CLIP (image)
- Encode captions with CLIP (text)
- Compute cross-modal similarity
- Keep candidates above threshold (semantic sanity check)

Step 3: Final Retrieval (Image-based)

- Compare ROI embedding with candidate rendered views (DINOv2)
- Select the model with the highest visual similarity

Object Retrieval Results

Datasets:

- YCB-V (21 objects), HouseCat6D (194 objects), YCB-V + GSO (1030 distractors)

Metrics:

- **mAP@k** → retrieval accuracy
- **Earth Mover's Distance (EMD)** → geometric similarity

	YCB-V	HCat6D	YCB-V&GSO
AP (%)	90.78	49.11	60.82
EMD	0.005	0.027	0.030

Table 1. Average EMD and AP across datasets

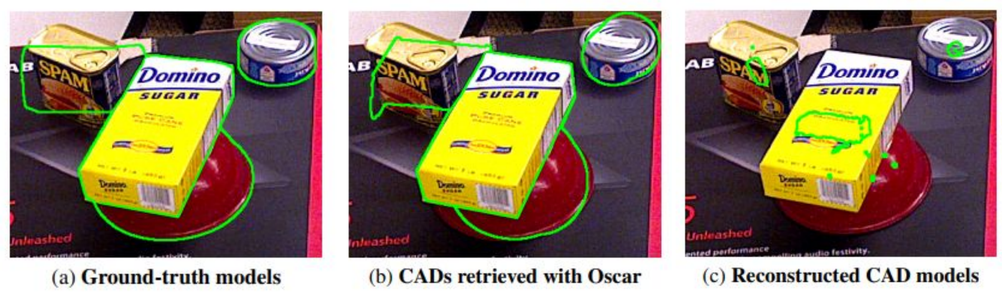
⇨ High AP on YCB-V, lower AP on larger/ diverse sets (HCat6D: 49.1%)

⇨ EMD remains close to 0 → retrieved models are still geometrically similar

Pose Estimation Results

Setup

- Evaluated MegaPose with three CAD model sources:
  - **Ground-truth** models (baseline)
  - Most-frequently **confused models** retrieved by OSCAR
  - **Reconstructed** models (Nerfacto, 50 images)
- Dataset: **YCB-V** test scene



(a) Ground-truth models (b) CADs retrieved with Oscar (c) Reconstructed CAD models

Object	Trl. error [mm]			Rot. error [°]		
	GT	OSCAR	Reconst.	GT	OSCAR	Reconst.
Sugar	16.29	20.19	2577.20	42.44	4.30	42.44
Tuna	29.53	806.40	5814.08	22.76	51.17	22.76
Bowl	25.06	85.55	3006.89	99.83	115.23	99.83
Spam	185.26	665.33	4282.35	87.83	57.17	87.83

Results

- **OSCAR**-retrieved models → more accurate poses than reconstruction-based models
- **Reconstruction** → consistently failed, produced unreliable poses
- **Ground-truth** models → best results overall, but some objects remain challenging (e.g., canned meat)

⇨ Even if OSCAR retrieves a confused but similar CAD model, MegaPose can still produce usable 6D poses.