JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

PartSLIP: Low-Shot Part Segmentation for 3D Point Clouds via Pretrained Image-Language Models

Minghua Liu[1] Yinhao Zhu[2] Hong Cai[2] Shizhong Han[2] Zhan Ling[1] Fatih Porikli[2] Hao Su[1]
[1]UC San Diego [2]Qualcomm AI Research

THU-PM-108

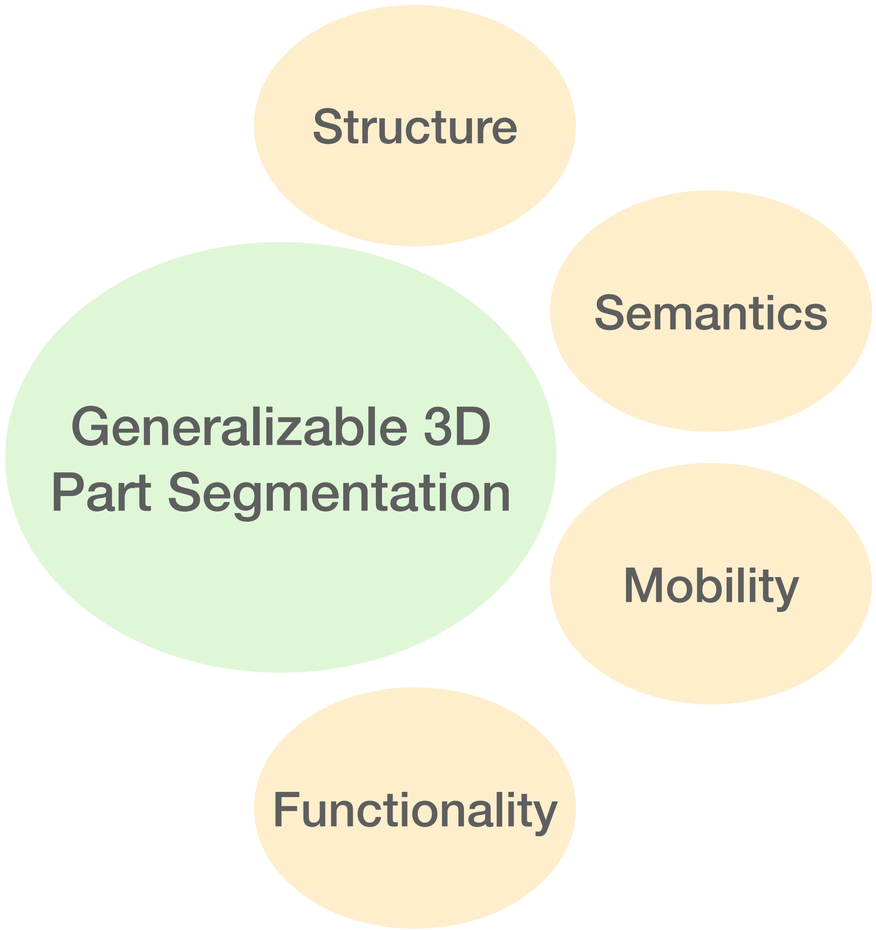
TL; DR

Zero/few-shot 3D part segmentation.

Highly competitive results compared to the fully supervised methods.

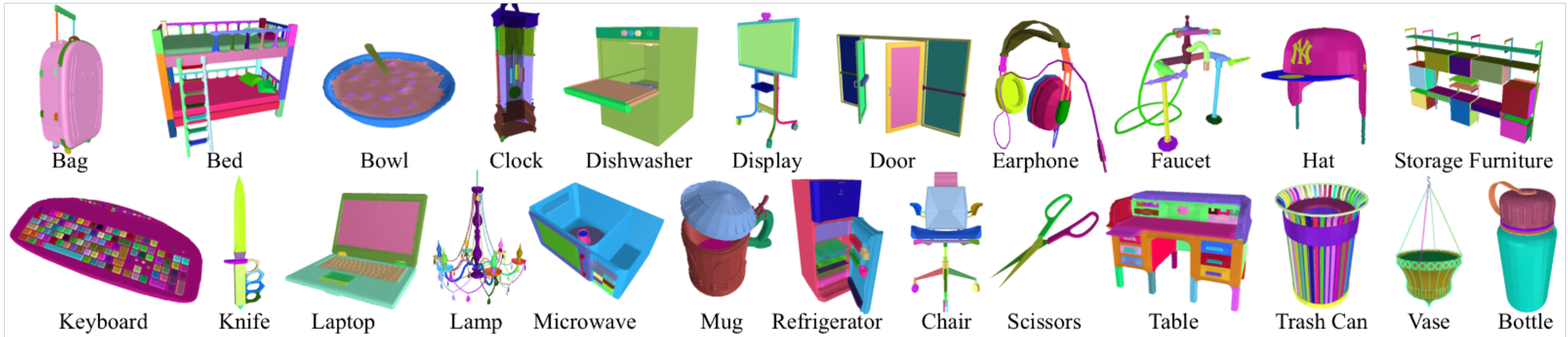
Can be directly applied to real-world point clouds without significant domain gaps.

3D Part Segmentation



Existing Approaches

- Supervised by 3D ground truth labels.
- Suffer from 3D (labeled) data scarcity.
  - E.g., PartNet only covers 24 object categories.
  - Poor generalization to unseen categories.

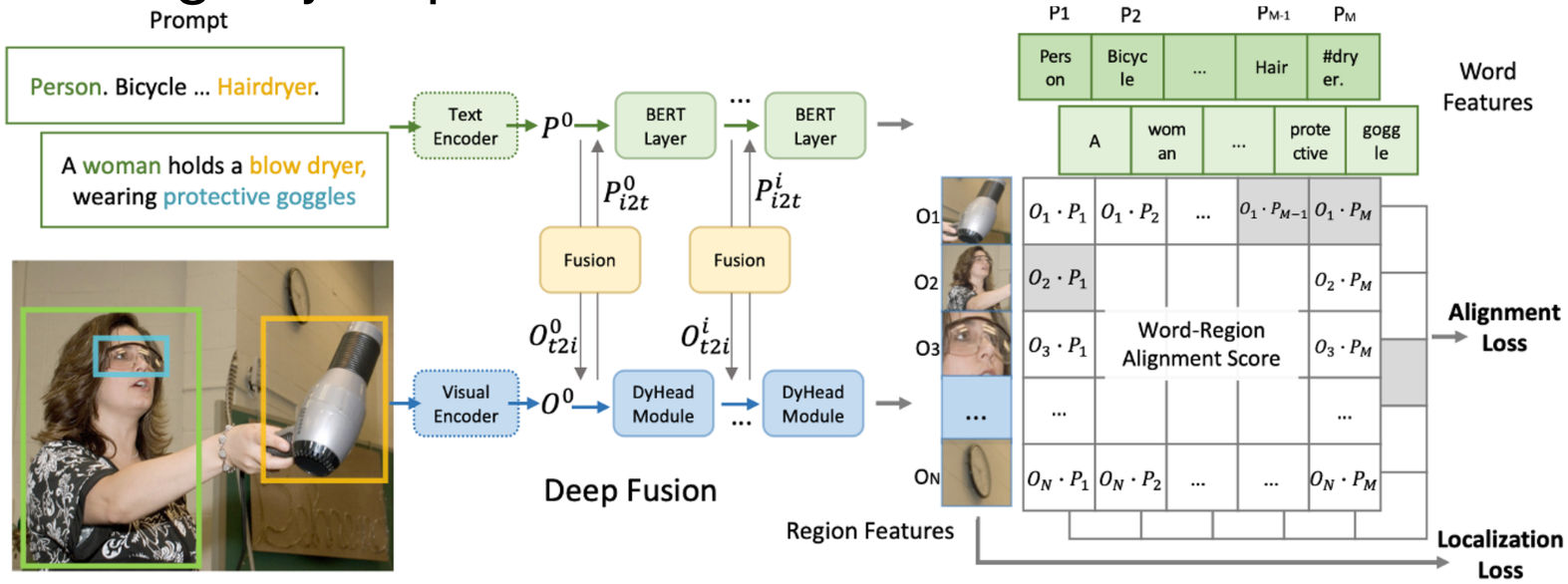


Mo, Kaichun, et al. "Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object".

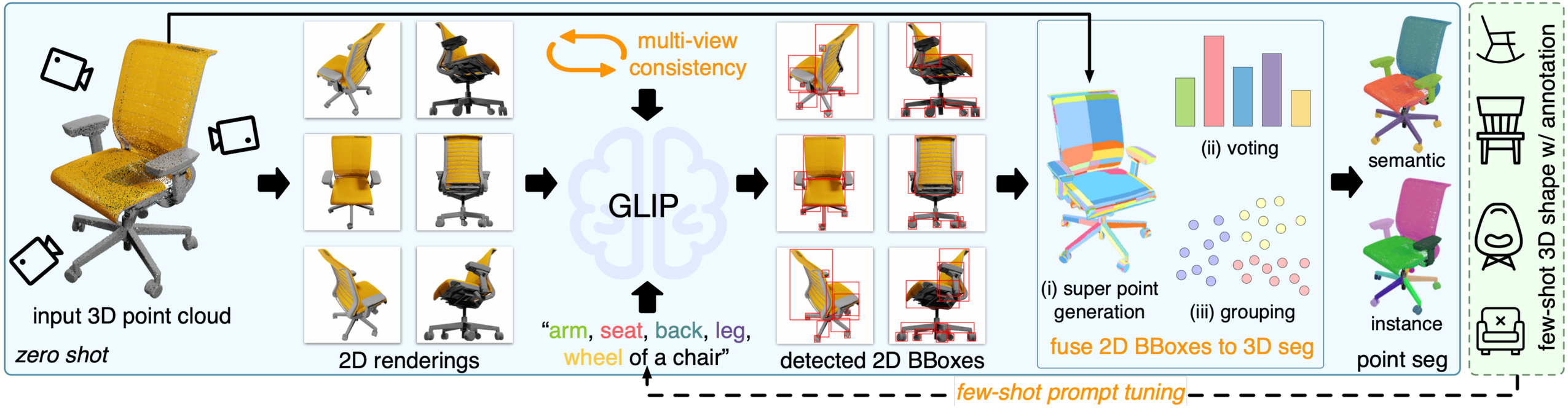
Leverage Pretrained Image-Language Models

GLIP

- Open-vocabulary 2D detection / grounding.
- Input: A free-form text description + a 2D image.
- Output: 2D bounding boxes.
- Excel at detecting object parts.

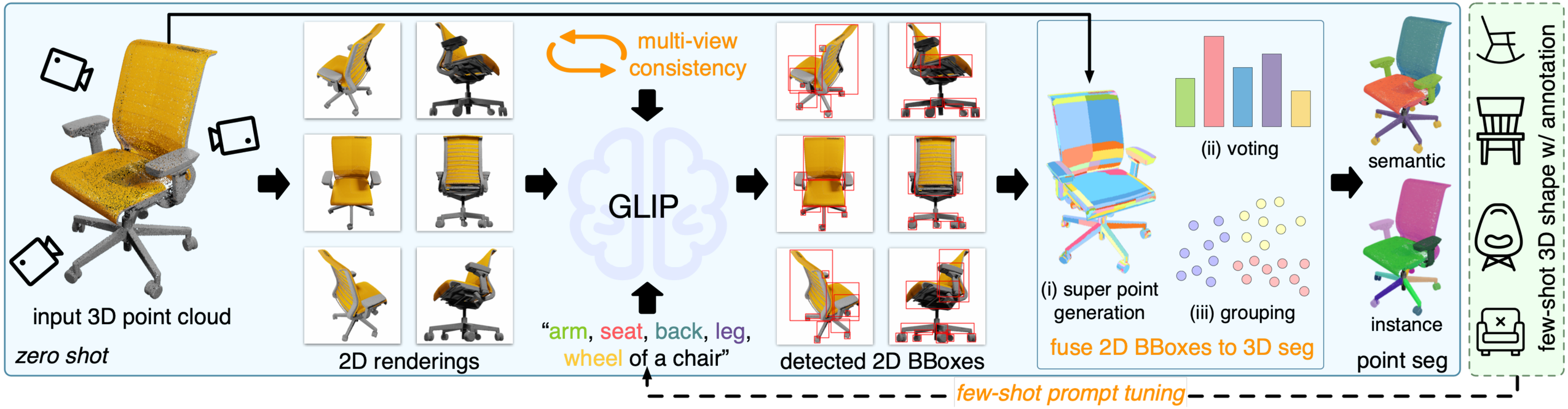


Pipeline


input 3D point cloud
zero shot
2D renderings
multi-view consistency
GLIP
"arm, seat, back, leg, wheel of a chair"
detected 2D BBoxes
(i) super point generation
(ii) voting
(iii) grouping
fuse 2D BBoxes to 3D seg
semantic
instance
point seg
few-shot 3D shape w/ annotation
few-shot prompt tuning

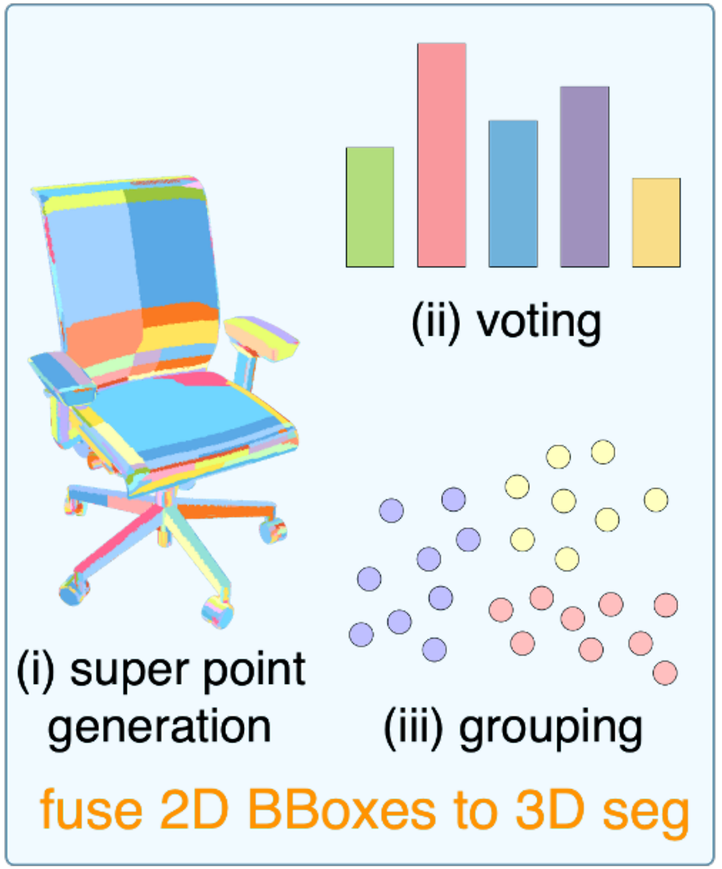
Pipeline

- How to convert 2D bboxes to 3D (semantic & instance) segmentation?
- How to finetune the GLIP model given few-shot 3D data?
- Can we leverage multi-view priors to boost GLIP's performance?



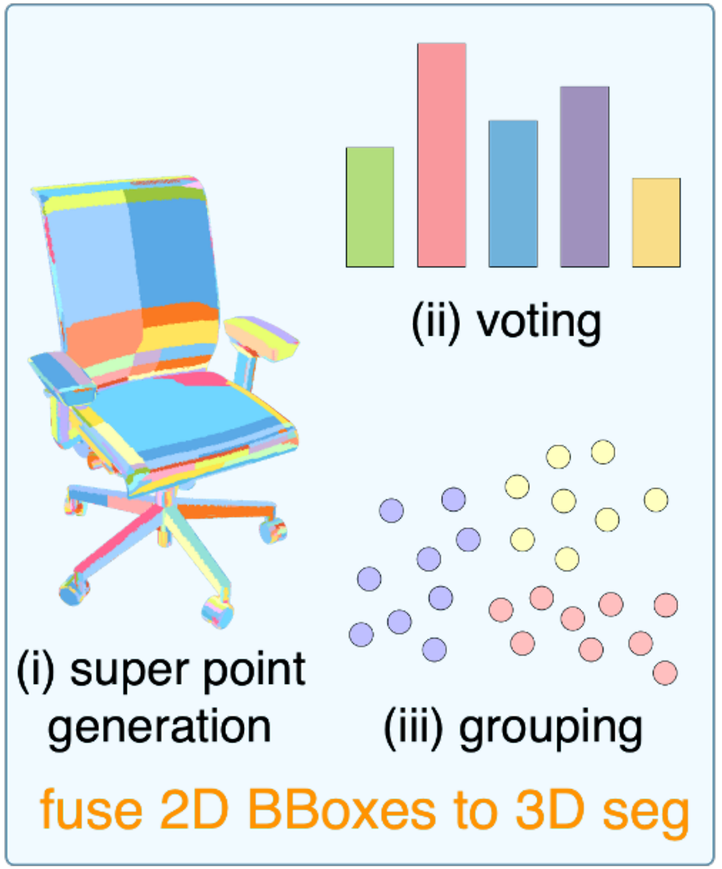
Detected 2D BBoxes to 3D Point Segmentation

- Challenges:
  - Bounding boxes are not as precise as point-wise labels.
  - Non-trivial to determine which sets of 2D bounding boxes indicate the same 3D part instance.
- A learning-free module:
  1. 3D super point generation
  2. 3D semantic voting
  3. 3D instance grouping



Detected 2D BBoxes to 3D Point Segmentation

1. 3D super point generation
   - Oversegment the input 3D point cloud into a collection of super points.
2. 3D semantic voting
   - Assign a semantic label for each super point.
3. 3D instance grouping
   - Group super points within each part category into instances based on their similarity of bounding box coverage.



Zero-Shot Segmentation

- Enable zero-shot open-vocabulary 3D part segmentation.
- Limited by GLIP's performances.

"door, handle of a refrigerator"

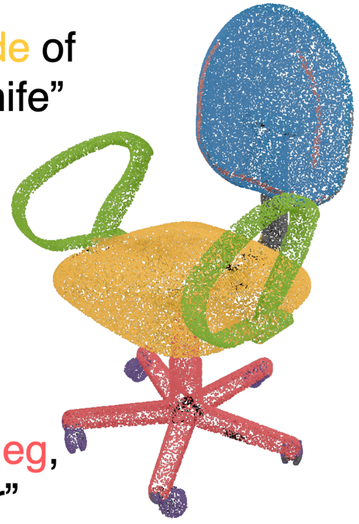
"cap of a pen"

"hand of a clock"

"handle, wheel of a suitcase"

"wheel of a cart"

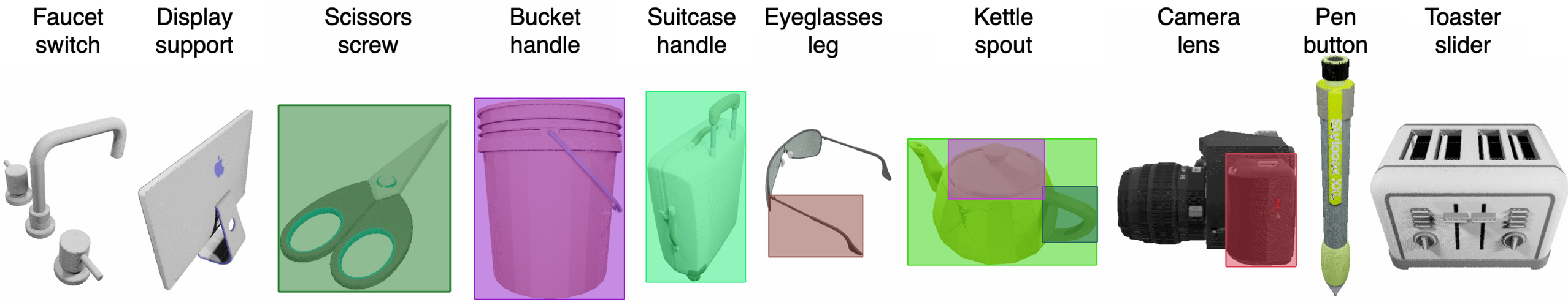
"blade of a knife"



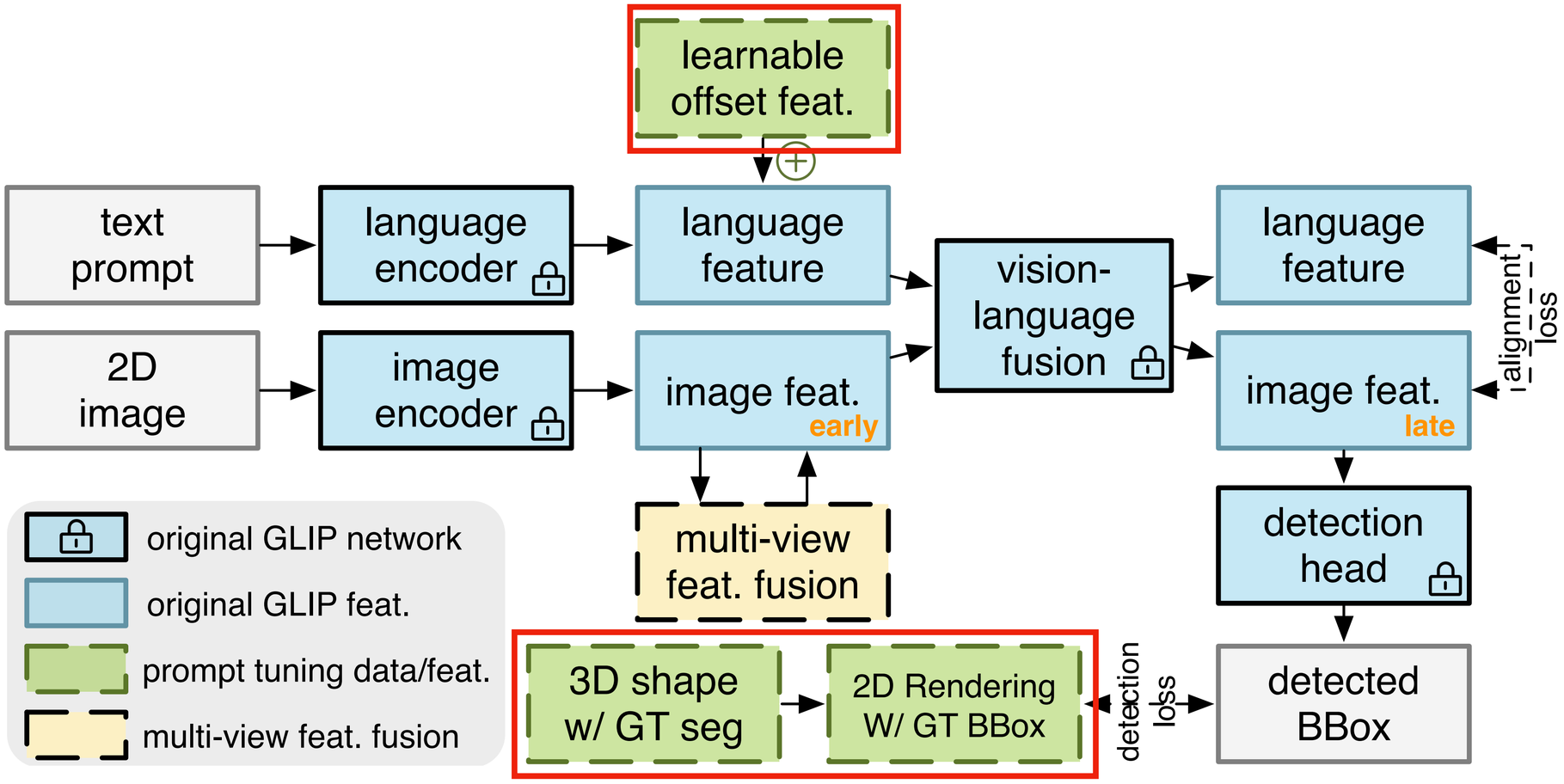
"back, arm, seat, leg, wheel of a chair"

GLIP Failure Cases

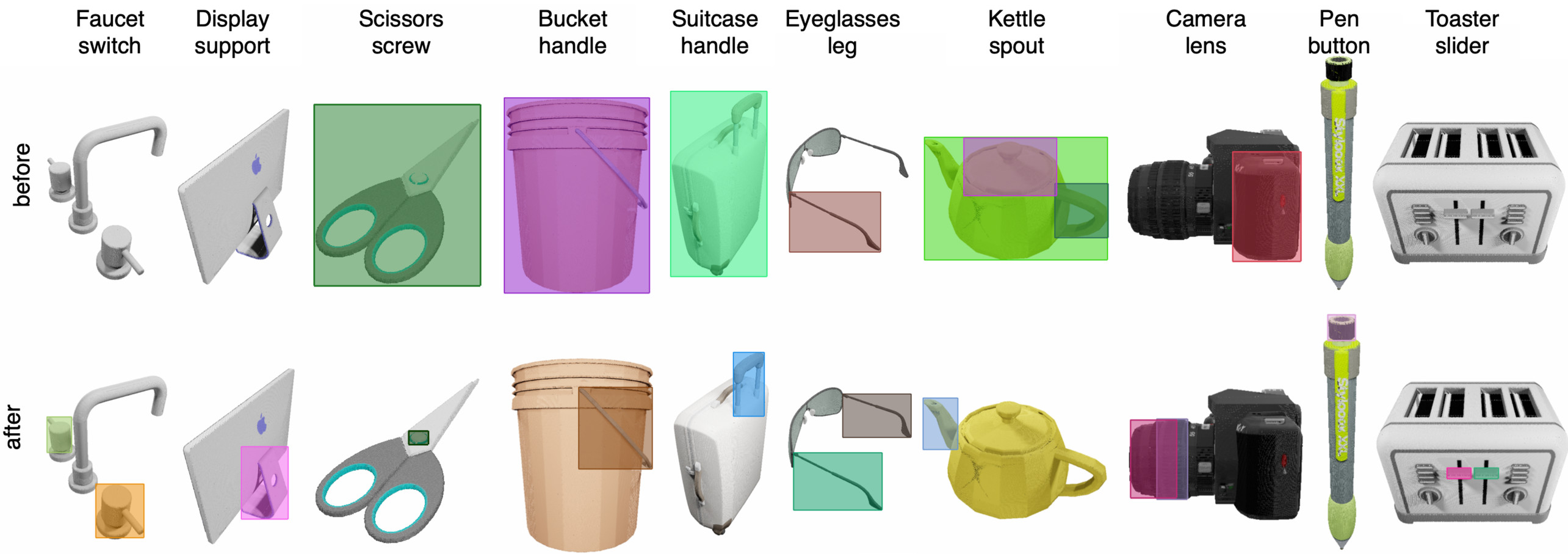
- Pretrained GLIP fail to understand some of our part definitions.
- Can we finetune GLIP model with a few 3D shapes with ground truth segmentation?



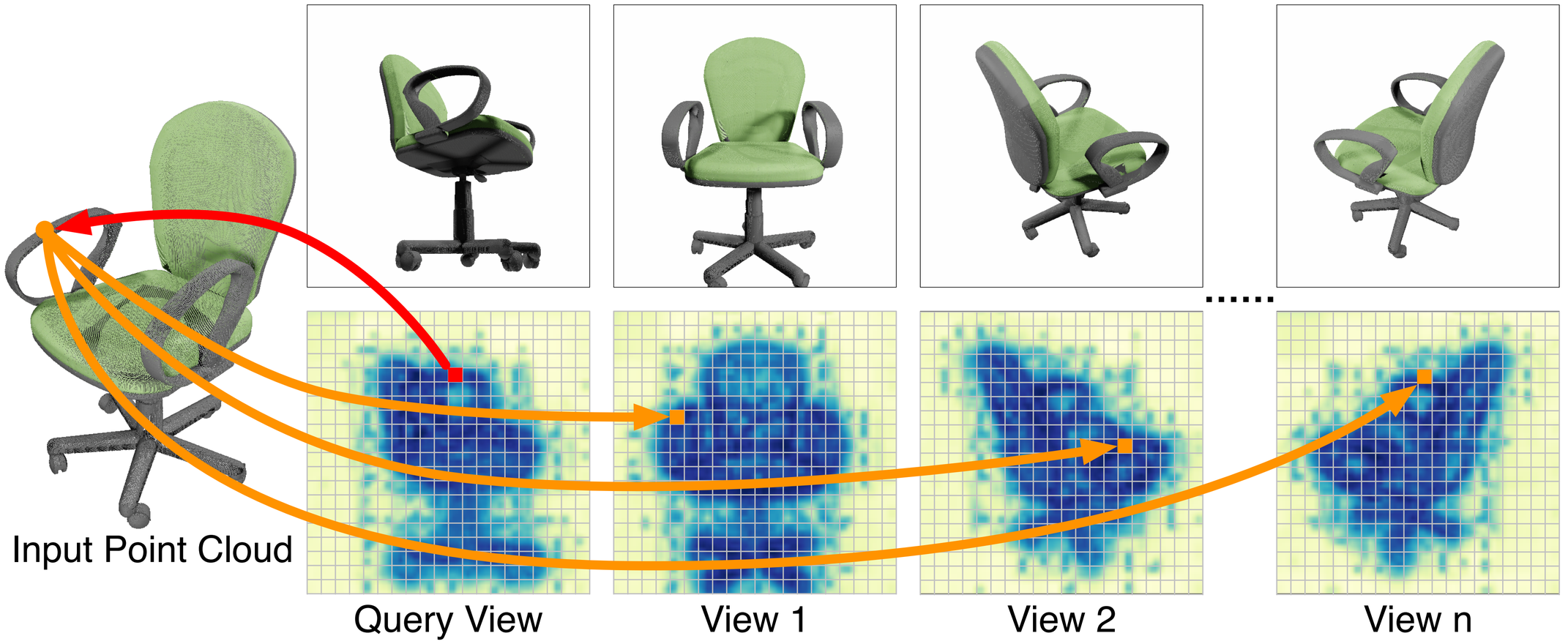
Few-Shot Prompt Tuning



Few-Shot Prompt Tuning

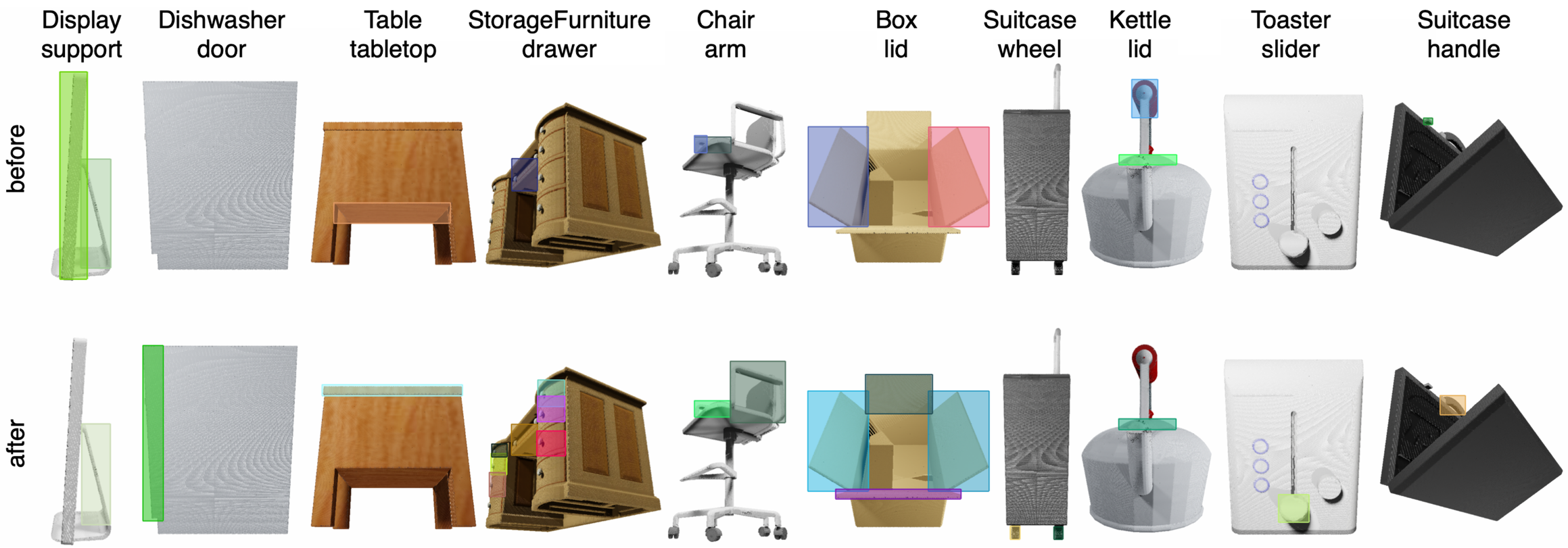


Multi-View Feature Aggregation

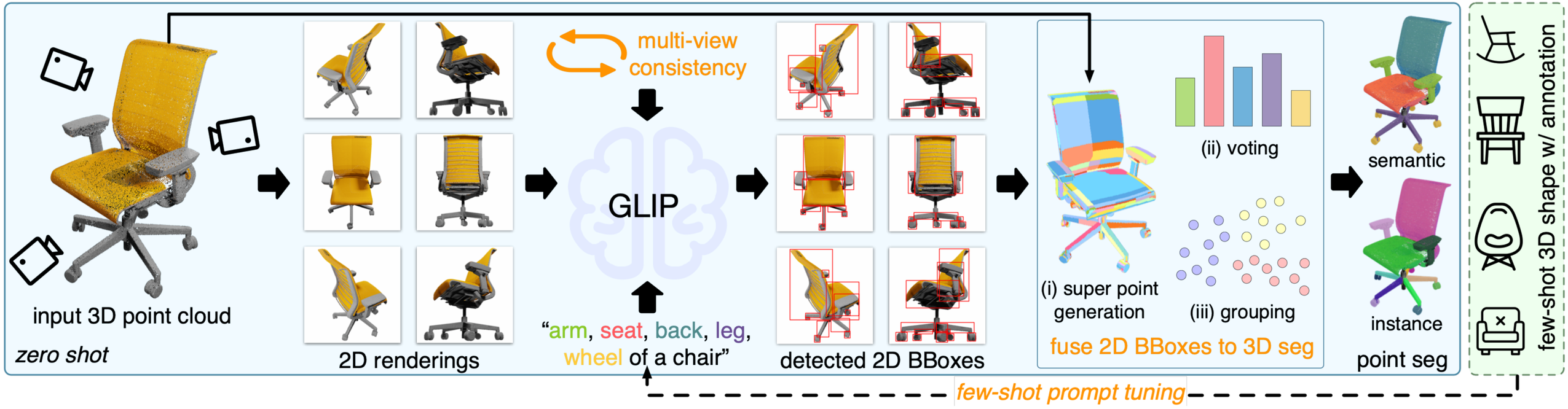


Multi-View Feature Aggregation

- Better handle images taken from some rare viewpoints.



Method


input 3D point cloud
zero shot
2D renderings
multi-view consistency
GLIP
"arm, seat, back, leg, wheel of a chair"
detected 2D BBoxes
(i) super point generation
(ii) voting
(iii) grouping
fuse 2D BBoxes to 3D seg
semantic
instance
point seg
few-shot 3D shape w/ annotation
few-shot prompt tuning

PartNet-Ensembled

- 45 object categories, 103 parts.

category	parts	few-shot	test	extra-train	category	parts	few-shot	test	extra-train
Bottle	lid	8	49	471	**Microwave**	display, door, handle, button	8	8	234
Box	lid	8	20	0	Mouse	button, cord, wheel	8	6	0
Bucket	handle	8	28	0	Oven	door, knob	8	22	0
Camera	button, lens	8	29	0	Pen	cap, button	8	40	0
Cart	wheel	8	53	0	Phone	lid, button	8	10	0
Chair	arm, back, leg, seat, wheel	8	73	8000	Pliers	leg	8	17	0
Clock	hand	8	23	593	Printer	button	8	21	0
CoffeeMachine	button, container, knob, lid	8	46	0	**Refrigerator**	door, handle	8	36	195
Dishwasher	door, handle	8	40	179	Remote	button	8	41	0
Dispenser	head, lid	8	49	0	Safe	door, switch, button	8	22	0
Display	base, screen, support	8	29	954	**Scissors**	blade, handle, screw	8	39	60
Door	frame, door, handle	8	28	237	Stapler	body, lid	8	15	0
Eyeglasses	body, leg	8	57	0	**StorageFurniture**	door, drawer, handle	8	338	2260
Faucet	spout, switch	8	76	681	Suitcase	handle, wheel	8	16	0
FoldingChair	seat	8	18	0	Switch	switch	8	62	0
Globe	sphere	8	53	0	**Table**	door, drawer, leg, tabletop, wheel, handle	8	93	9799
Kettle	lid, handle, spout	8	21	0	Toaster	button, slider	8	17	0
Keyboard	cord, key	8	29	165	Toilet	lid, seat, button	8	61	0
KitchenPot	lid, handle	8	17	0	**TrashCan**	footpedal, lid, door	8	62	358
Knife	blade	8	36	505	USB	cap, rotation	8	43	0
Lamp	base, body, bulb, shade	8	37	3246	WashingMachine	door, button	8	9	0
Laptop	keyboard, screen, shaft, touchpad, camera	8	47	430	Window	window	8	50	0
Lighter	lid, wheel, button	8	20	0	**45 in total**	**103 in total**	**360**	**1,906**	**28,367**

Quantitative Results

- Impressive zero-shot performances.
- Not only outperforms existing few-shot approaches by a large margin, but also highly competitive compared to the fully supervised counterparts.

Semantic Segmentation

#3D data	Method	mIoU
few-shot w/ extra data (45x8 + 28k)	PointNet++	36.8
	PointNeXt	**50.2**
	SoftGroup	38.1
few-shot (45x8)	PointNet++	20.4
	PointNeXt	40.6
	SoftGroup	38.0
	ACD	23.2
	Prototype	44..3
	PartSLIP	**59.4**
zero-shot	PartSLIP	**34.8**

Instance Segmentation

#3D data	Method	mAP50
few-shot w/ extra data (45x8 + 28k)	PointGroup	31.0
	SoftGroup	**31.9**
few-shot (45x8)	PointGroup	16.0
	SoftGroup	25.7
	PartSLIP	**44.8**
zero-shot	PartSLIP	**18.0**

Real-World Point Clouds

- Input point clouds scanned by an iPhone.

Takeaways

- A novel approach for low-shot 3D part segmentation.
- Achieves impressive zero-shot performances and highly competitive few-shot results compared to the fully supervised counterparts.
- Can be applied to real-world point clouds without significant domain gap.