ObjVariantEnsemble: Advancing Point Cloud LLM Evaluation in Challenging Scenes with Subtly Distinguished Objects

Anonymous submission

1 Technical Appendix

In this section, we offer further explanations about the construction of ObjVariantEnsemble . Section 1.1 details the annotation process, focusing on the algorithm used. Section 1.2 outlines the instruction prompts employed for LLMs and VLMs. Finally, Section 1.3 presents additional experimental results from our evaluation of the object-level model.

1.1 Annotation

This section offers additional details on how to generate descriptions that highlight distinctions, aiding in the discrimination between targets and distractors.
The algorithm we used in this process is detailed in Algorithm1 and Algorithm2.**The related code will be released upon acceptance.** The following variables are used in the algorithm:

- **target**: The target object in the scene.
- **candidate**: The candidate object selected as distractors.
- **round**: Represents each round within the loop structure.
- **view**: Refers to different viewpoints or perspectives of the objects.
- **target_descriptions**: A collection of descriptions generated for the target object.
- **candidate_descriptions**: A collection of descriptions generated for the candidate object.
- **image**: The concatenated image that combines views of the target and candidate objects.
- **difference**: The difference captured from the concated image.
- **difference_all**: A collection of all differences from different views.
- **summary**: The summary of all captured differences between the target and candidate.
- **target_view**: Rendered images from a single view of the target object.
- **candidate_view**: Rendered images from a single view of the candidate object.
- **target_descriptions**: The summarized description for the target object.
- **candidate_descriptions**: The summarized description for the candidate object.
- **answers_all**: A list that stores all question-answer pairs generated during the iterative process.
- **q_next**: The next question generated in each iteration, based on the accumulated answers and the given instructions.
- **a_next**: The next answer generated in each iteration, based on the accumulated answers and the given instructions.
- **key_difference**: The final summary generated by the GPT model that highlights the most important distinctions between the target and candidate objects.
- **Annotation**: The final annotation generated based on the target summary, candidate summary, and overall summary.
- **FIRST_QUESTION**: The initial question used to generate the first response about the image. This serves as the starting point for the iterative caption generation process.
- **QUESTION_INSTRUCTION**: The instruction provided to the GPT model to guide the generation of the next question based on the accumulated answers.
- **SUB_QUESTION_INSTRUCTION**: Additional instructions that help the GPT model focus on generating sub-questions that refine or clarify specific details in the image.
- **SUMMARY_INSTRUCTION_1**: The instruction given to the GPT model to generate a summary that captures the key differences between the target and candidate objects based on all the gathered question-answer pairs.
- **SUMMARY_INSTRUCTION_2**: Prompt provided to the GPT model for summarizing descriptions from multiple views.
- **SUMMARY_INSTRUCTION_3**: Prompt provided to the GPT model for generating the final annotation.

Spatial Relationship Enhancement For a complete annotation, we further enhance the spatial location information based on the integration specifications used for scene construction. We curated a vocabulary dictionary for spatial primitives and their combinations, which is shown in Table3.

1.2 Prompt

The prompts used in the annotation process are shown in Figure 1. These correspond to the variables mentioned earlier, including FIRST_QUESTION, QUESTION_INSTRUCTION, and SUB_QUESTION_INSTRUCTION for LLaVA, as well as SUMMARY_INSTRUCTION_1, SUMMARY_INSTRUCTION_2, and SUMMARY_INSTRUCTION_3 for GPT-3.5-turbo.

Algorithm 1: Distinction Annotating Process

for each (target, candidate) in pairs **do**
  for each round in rounds **do**
    for each view in views **do**
      target_descriptions.append(view: LLaVA(target_view))
      candidate_descriptions.append(view: LLaVA(candidate_view))
      image ← Concat(target_view, candidate_view)
      difference ← Iter_cap(image, iter_rounds) {Difference capturing}
      difference_all.append(view:difference)
    end for
    summary ← GPT(difference_all, SUMMARY_INSTRUCTION_2)
    target_description ← GPT(target_descriptions, SUMMARY_INSTRUCTION_2)
    candidate_description ← GPT(candidate_descriptions, SUMMARY_INSTRUCTION_2)
  end for
  Annotation ← GPT(target_description, candidate_description, summary, SUMMARY_INSTRUCTION_3)
end for

Algorithm 2: Iterative Caption Generation (Iter_cap)

Require: image, iter_rounds, FIRST_QUESTION, QUESTION_INSTRUCTION, SUB_QUESTION_INSTRUCTION, SUMMARY_INSTRUCTION_1
first_q ← FIRST_QUESTION
first_a ← LLaVA(image, first_q)
answers_all ← [first_q, first_a]
for each iter in iter_rounds **do**
  q_next ← GPT(answers_all, QUESTION_INSTRUCTION, SUB_QUESTION_INSTRUCTION)
  a_next ← LLaVA(image, q_next)
  answers_all.append([q_next, a_next])
end for
key_difference ← GPT(answers_all, SUMMARY_INSTRUCTION_1)
return key_difference

Model	Average.	Airplane	Bag	Cap	Car	Chair	Earphone	Guitar	Knife	Lamp	Laptop	Motorbike	Mug	Pistol	Rocket	Skateboard	Table
PointBert(sc)	89.43	85.62	**49.50**	71.39	80.68	89.98	67.78	89.38	81.43	88.45	94.70	60.15	96.03	91.53	78.03	89.43	92.17
Uni3d(sc)	80.05	79.80	**25.98**	85.56	71.23	84.82	91.41	73.48	85.67	75.41	84.91	45.52	94.41	87.75	77.31	71.54	84.42
PointBert(dc)	79.90	79.08	88.21	91.44	83.14	83.01	71.28	94.16	75.51	**70.12**	95.77	78.13	81.35	85.19	86.31	92.99	78.74
Uni3d(dc)	66.04	61.94	62.03	87.68	82.87	83.03	72.88	87.80	79.42	62.88	94.22	77.94	86.68	75.44	61.74	90.87	**46.72**

Table 1: Segmentation evaluation(Loc+Shape) on PointBERT(Yu et al. 2022). Models trained on scene data constructed from the different class tend to confuse objects of different types.

1.3 Experimental Results

Here we offer more segmentation results for each category. For our evaluation, we selected Uni3D(Zhang et al. 2023) and PointBERT(Yu et al. 2022), which represent state-of-the-art performance. However, considering the data adaptation issue, where models may need fine-tuning to perform well on new tasks, we added extra fully connected layers to fine-tune existing models on the tasks we constructed. We used two types of data for fine-tuning: 'sc' refers to scenes containing objects from the same class, while 'dc' refers to scenes with objects from different classes. The performance metric used was mIoU (%), which stands for mean Intersection over Union. The results from Table 1 and Table 2 provide further evidence that models trained on scene data constructed with the same class (more similar distractors) demonstrate better performance.

Model	Average.	Round with handle	Rectangle	L-shaped	Wheeled
PointBert(sc)	77.29	73.87	88.26	93.86	**53.17**
Uni3d(sc)	**65.09**	**69.38**	**50.69**	86.14	54.79
PointBert(dc)	76.44	92.41	54.94	**84.02**	74.39
Uni3d(dc)	74.85	69.84	71.65	87.69	70.23

Table 2: Segmentation evaluation(Loc+Class) on PointBERT(Yu et al. 2022). Models trained on scene data constructed from the same class tend to confuse objects with similar shapes. The overall performance of models trained on same-class data remains comparable to those trained on data from different classes, with PointBert(sc) even performing better in some cases.

Functions	Vocabularies
Left	On the left of/ To the left side of/ Leftward of...
Right	On the right of/ To the right side of/ Rightward of...
Front	Ahead of/ In front of/ Before...
Back	Behind/ At the back of/ After...
Up	Above/ On top of/ At the top/ Over...
Down	Below/ Underneath/ Bottom of/ Under...
Left+Up	Upper left/ Top left...
Right+Up	Upper right/ Top right...
Left+Down	Lower left/ Bottom left...
Right+Down	Lower right/ Bottom right...
Left+Right/Front+Back/Top+Down	Between/ Flanked by/ In line with...
Left+Right+Front+Back	Surrounding by/ Among/ Enclosed by...
Left/Right/Front/Back/Up/Down	Near/ Close to/ Adjacent to...

Table 3: Vocabulary dictionary for spatial primitives and their combinations.

DIFFERENCE_CAPTURING_INSTRUCTION *(for LLaVA)*

FIRST_QUESTION

"Descript the 3D object in the photo."

QUESTION INSTRUCTION

"I have an image with two objects placed on the left and right."
"Ask me questions about the content of this image. "
"Carefully asking me informative questions to maximize your information about difference of these two objects."
"Each time ask one question only without giving an answer. "
"Avoid asking yes/no questions."
"Ask more information about difference in shape. e.g. The first anise has a more compact outline and a less symmetrical shape, while the second anise has a broader outline and a more symmetrical shape"
"Avoid asking information about color and texture."
"Avoid asking information about background."
'I\'ll put my answer beginning with "Answer:".'

SUB_QUESTION_INSTRUCTION

"Next Question. Avoid asking yes/no questions. \n" "Question:"

SUMMARY_INSTRUCTION_1 *(for Gpt3.5-turbo)*

"Now summarize the information you get in a few words.\n
Don't add the not sure or negative information into summary. \n
Remember you do not need to summary all sentences. Ignore the sentences with answers negative or not sure.\n
Don't imagine or add information.\n
Don't add negative statement or not sure statement into summary.\n
Summary: Left object is ..., right object is ... "

SUMMARY_INSTRUCTION_2 *(for Gpt3.5-turbo)*

Given discription of {category} from three viewpoints: {Multi_view descriptions}.\n\
please summarize the description in one sentence \n\
Avoid uncertain or negative information. Avoid describing the background.

SUMMARY_INSTRUCTION_3 *(for Gpt3.5-turbo)*

"Description of left {category}: {caption_obj_left} \n\
Description of right {category}:{caption_obj_left}\n \
Given difference between left {category} and right {category}: {summary1}. \n\
Summarize the description for right {category} in one sentence. \n\n\ "

"Avoid uncertain or negative information. Avoid describing the background. \n\
Don't add the not sure or negative information into summary. \n
Don't imagine or add information.\n
Don't include words like left and right."

Figure 1: Prompts used in difference capturing and summarizing. DIFFERENCE_CAPTURING_INSTRUCTION is designed for LLaVA to capture differences between targets and candidates from multi-view images. SUMMARY_INSTRUCTION_1 is designed for Gpt3.5_turbo to summarize multi-round descriptions of differences. SUMMARY_INSTRUCTION_2 is designed for Gpt3.5_turbo to summarize multi-view differences. SUMMARY_INSTRUCTION_3 is designed for Gpt3.5_turbo to enhance annotations with captions of targets and differences with candidates.

2 Dataset Appendix

This section provides additional visualizations of our dataset along with statistics from the construction process. Section2.1 presents statistics on the number of object categories included in ObjVariantEnsemble from both scene-level and object-level datasets. Section2.1 offers further data visualizations in ObjectVariantEnsemble.

2.1 Data statistics

In the process of searching for potential candidates in object-level datasets, each category yielded both successful and unsuccessful cases. We used NYU40ID to segment and classify each object in ScanNet, organizing them into background and target object lists. For each category in the target object list, we searched for similar objects within the same category or with a similar shape. The words in red in Figure 4 indicate the objects for which candidates were successfully found. Table 4 provides examples of the number of object types identified in the object-level datasets(Chang et al. 2015; Uy et al. 2019; Sun et al. 2022; Wu et al. 2023).

Category	Object-level dataset
Cabinet	Omniobject(9), Shapenet(1571), Scanobjectnn(347)
Chair	Omniobject(29), Modelnet40(989), Shapenet(6778), Scanobjectnn(395)
Table	Omniobject(28), Modelnet40(492), Shapenet(8434), Scanobjectnn(241)
Box	Omniobject(136), Scanobjectnn(117)
Stool	Omniobject(11), Modelnet40(110)
...	...

Table 4: Potential object types in the object-level datasets.

2.2 Data visualization

We provide additional data visualizations in ObjectVariantEnsemble, covering various categories and tasks. Samples of our data are included in the .zip file submitted with our materials.

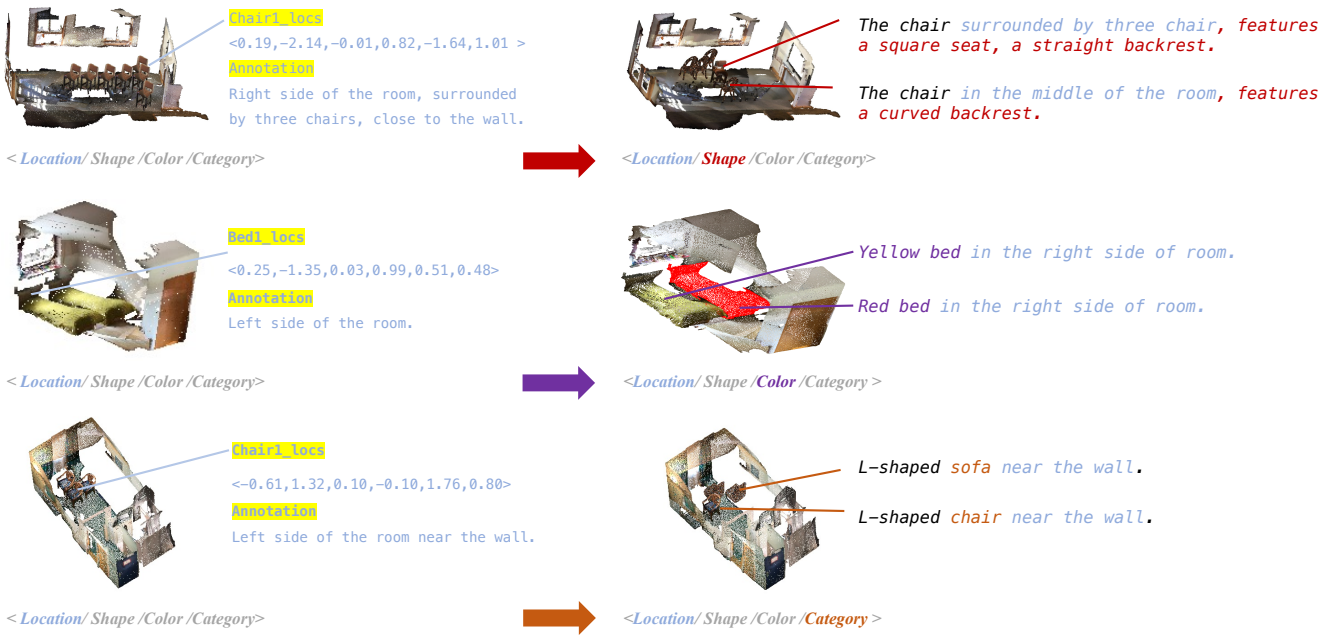

Figure 2: Visualizations of our tasks.

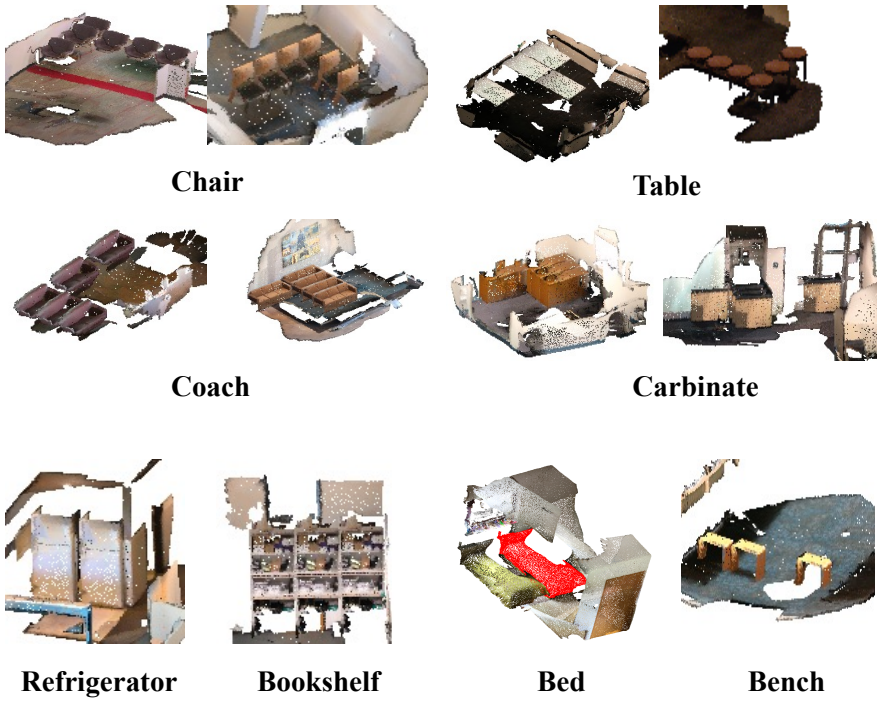

Figure 3: Visualizations of our datasets.

Label index	*Object categories*
1	wall, shower walls, closet wall, shower wall, pantry wall, closet walls, bath walls, pantry walls, door wall
2	floor, shower floor, closet floor
3	cabinet, kitchen cabinet, kitchen cabinets, file cabinet, bathroom vanity, cabinets, bathroom cabinet, cabinet doors, open kitchen cabinet, file cabinets, trash cabinet, media center
4	bed, mattress, loft bed, sofa bed, air mattress
5	chair, office chair, armchair, sofa chair, stack of chairs, folded chair, folded chairs, massage chair, recliner chair, rocking chair, stack of folded chairs
6	couch, sofa
7	table, coffee table, end table, dining table, folded table, round table, side table, air hockey table
8	door, doorframe, doors, bathroom stall door, closet doors, closet door, shower door, mirror doors, cabinet door, glass doors, sliding door, closet doorframe
9	window
10	bookshelf, bookshelves
11	picture, poster, painting, pictures
12	kitchen counter, counter, bathroom counter
13	blinds
14	desk
15	shelf, organizer shelf, pantry shelf, closet shelf
16	curtain, curtains
17	dresser
18	pillow, pillows, couch cushions, cushion
19	mirror
20	mat, yoga mat
21	clothes, clothing, cloth, sock, kitchen apron, costume, sock
22	ceiling, closet ceiling
23	books, book, music book
24	refrigerator, mini fridge, cooler
25	tv
26	paper2, papers
27	towel, towels, hand towel
28	shower curtain
29	box, boxes, mailboxes, mailbox, storage box, pizza box, boxes of paper, jewelry box, cat litter box, covered box, pizza boxes
30	whiteboard
31	person, legs
32	nightstand
33	toilet, urinal
34	sink
35	lamp, lamp base, desk lamp, wall lamp, table lamp, ceiling lamp, night lamp
36	bathtub
37	bag, paper bag, messenger bag, ikea bag, duffel bag, bag of coffee beans, grocery bag, golf bag, garbage bag, coffee bean bag, trash bag, cosmetic bag, shopping bag, food bag
38	board, stove, light, bathroom stall, bar, light switch, ceiling light, range hood, blackboard, rail, bulletin board, ledge, shower, windowsill, dishwasher, stair rail, stairs, handicap bar, column, oven, pillar, structure, shower head, projector screen, staircase, fireplace, breakfast bar, hand rail, water fountain, kitchen island, pipes, shower control valve, handrail, step, dart board, grab bar, railing, stair, soap bar, studio light, shower doors, boards, frame, garage door, platform, elevator, wood beam, banister, curtain rod, chandelier, stovetop, glass
39	trash can, radiator, recycling bin, ottoman, bench, tv stand, wardrobe closet, trash bin, seat, closet, ladder, piano, water cooler, stand, washing machine, rack, washing machines, wardrobe cabinet, clothes dryer, ironing board, keyboard piano, music stand, furniture, crate, clothes dryers, drawer, footrest, piano bench, foosball table, footstool, compost bin, tripod, treadmill, chest, folded ladder, drying rack, pool table, heater, toolbox, beanbag chair, dollhouse, ping pong table, clothing rack, podium, luggage stand, rack stand, futon, book rack, seating, workbench, easel, luggage rack, headboard, display rack, crib, bedframe, closet wardrobe, wardrobe, bunk bed, magazine rack, furnace, stepladder, baby changing station, flower stand, display
40	object, monitor, backpack, plant, toilet paper, shoes, keyboard, bottle, stool, kettle, computer tower, telephone, cup, jacket, microwave, paper towel dispenser, suitcase, laptop, printer, soap dispenser, fan, tissue box, blanket, copier, soap dish, laundry hamper, storage bin, coffee maker, decoration, clock, mouse, basket, dumbbell, bucket, sign, speaker, container, shower curtain rod, tube, storage container, paper towel roll, ball, laundry basket, cart, dish rack, purse, bicycle, tray, plunger, paper cutter, toilet paper dispenser, bin, toilet seat cover dispenser, guitar, fire extinguisher, pipe, vacuum cleaner, plate, cd case, bowl, closet rod, scale, broom, hat, guitar case, water pitcher, laundry detergent, hair dryer, divider, power outlet, coffee kettle, toaster, shoe, alarm clock, water bottle, case of water bottles, toaster oven, coat rack, storage organizer, machine, fire alarm, vent, power strip, calendar, toilet paper holder, potted plant, stuffed animal, luggage, headphones, candle, projector, dustpan, rod, globe, step stool, vending machine, ceiling fan, swiffer, jar, hamper, poster tube, case, carpet, thermostat, coat, smoke detector, flip flops, banner, clothes hanger, whiteboard eraser, iron, instrument case, toilet paper rolls, soap, block, wall hanging, toothbrush, shirt, cutting board, vase, exercise machine, shorts, tire, teddy bear, bathrobe, faucet, thermos, rug, tupperware, shoe rack, beer bottles, salt, dispenser, remote, carton, slippers, soda stream, toilet brush, cooking pot, stapler, scanner, elliptical machine, kettle, metronome, dumbell, rice cooker, sewing machine, flowerpot, nerf gun, binders, quadcopter, pitcher, hanging, mail, hoverboard, water heater, spray bottle, rope, plastic container, soap bottle, sleeping bag, frying pan, oven mitt, pot, hand dryer, shampoo bottle, hair brush, tennis racket, display case, boiler, bananas, carseat, helmet, umbrella, coffee box, envelope, wet floor sign, controller, dolly, shampoo, paper tray, changing station, poster printer, screen, crutches, stack of cups, toilet flush button, trunk, plastic bin, car, shaving cream, shredder, statue, hose, bike pump, coatrack, bear, humidifier, toothpaste, mouthwash bottle, poster cutter, food container, camera, card, mug, cardboard, flag, magazine, exit sign, rolled poster, wheel, blackboard eraser, organizer, doll, laundry bag, sponge, lotion bottle, can, lunch box, food display, storage shelf, sliding wood door, pants, wood, bottles, washcloth, cups, exercise ball, roomba, bike lock, briefcase, bath products, star, map, ipad, traffic cone, toiletry, canopy, paper organizer, barricade, cap, dumbbell plates, cooking pan, santa, boat, kinect, plastic storage bin, dishwashing soap bottle, xbox controller, banana holder, ping pong paddle, airplane, conditioner bottle, tea kettle, toilet paper package, wall mounted coat rack, film light, chain, sweater, kitchen mixer, water softener, trolley, loofa, shower faucet handle, toy piano, fish, electric panel, suitcases, tape, plates, alarm, fire hose, toy dinosaur, cone, hatrack, subwoofer, fire sprinkler, photo, barrier, stacks of cups, beachball, folded boxes, contact lens solution bottle, folder, mail trays, slipper, sticker, lotion, buddha, file organizer, paper towel rolls, fuse box, knife block, cd cases, stools, hand sanitzer dispenser, teapot, pen holder, tray rack, wig, switch, plastic containers, night light, notepad, mail bin, elevator button, gaming wheel, drum set, coffee mug, baby mobile, diaper bin, stepstool, paper shredder, dress rack, cover, exercise bike, kitchenaid mixer, soda can, tap, cable, binder, towel rack, medal, telescope, baseball cap, battery disposal jar, mop, tank, mail tray, centerpiece, stick, dryer sheets, bycicle, clip, postcard, display sign, paper towel, boots, tennis racket bag, clothes hangers, starbucks cup

Figure 4: Object list categorized in semantic labels from NYU40ID.

References

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Sun, J.; Zhang, Q.; Kailkhura, B.; Yu, Z.; Xiao, C.; and Mao, Z. M. 2022. Benchmarking robustness of 3d point cloud recognition against common corruptions. *arXiv preprint arXiv:2201.12296*.

Uy, M. A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, T.; and Yeung, S.-K. 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1588–1597.

Wu, T.; Zhang, J.; Fu, X.; Wang, Y.; Ren, J.; Pan, L.; Wu, W.; Yang, L.; Wang, J.; Qian, C.; et al. 2023. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 803–814.

Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; and Lu, J. 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19313–19322.

Zhang, B.; Yuan, J.; Shi, B.; Chen, T.; Li, Y.; and Qiao, Y. 2023. Uni3d: A unified baseline for multi-dataset 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9253–9262.