3D Visual Prompting for Visual Question Answering

Members of Team 7 :

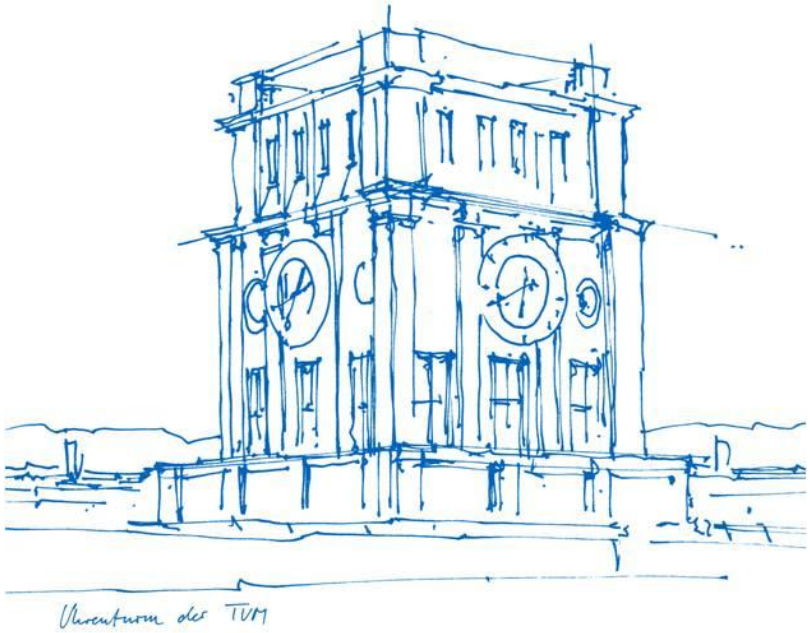
Bingning Huang

Nigar Doga Karacik

Ran Ding

Advisor: Mahdi Saleh

Client: Niko, Felix



Overview

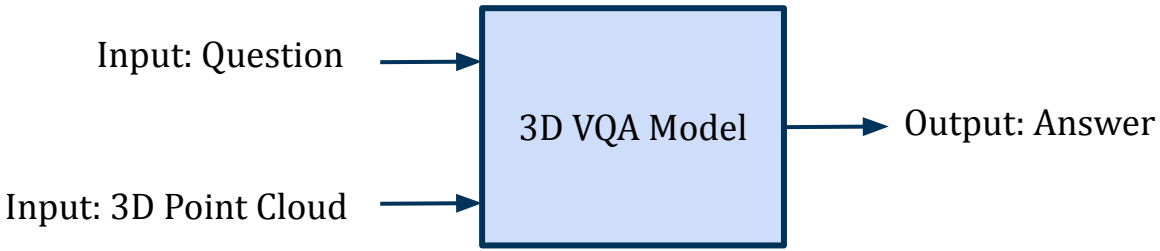
1. Motivation & Task statement
2. Related work
3. Methodology
4. Results
5. Future work
6. Lessons learned & Feedback on the course
7. Reference

1. Motivation & Task statement

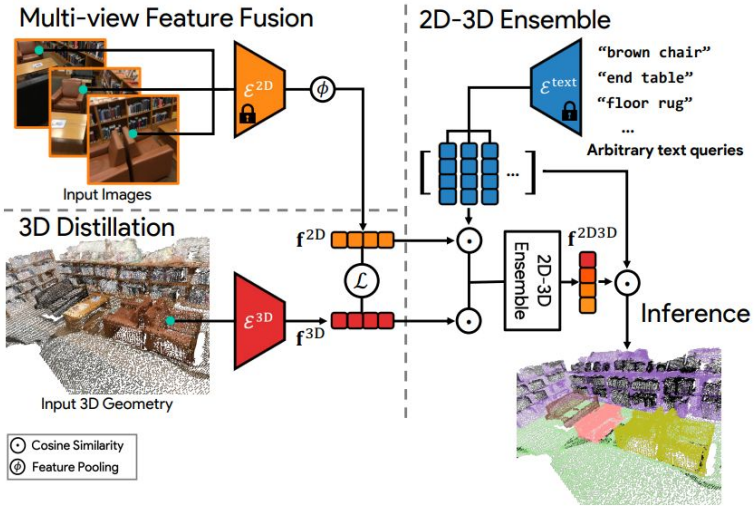
Task: Develop a model that can directly take 3D point cloud as input and perform Visual Question Answering task.

Applications:

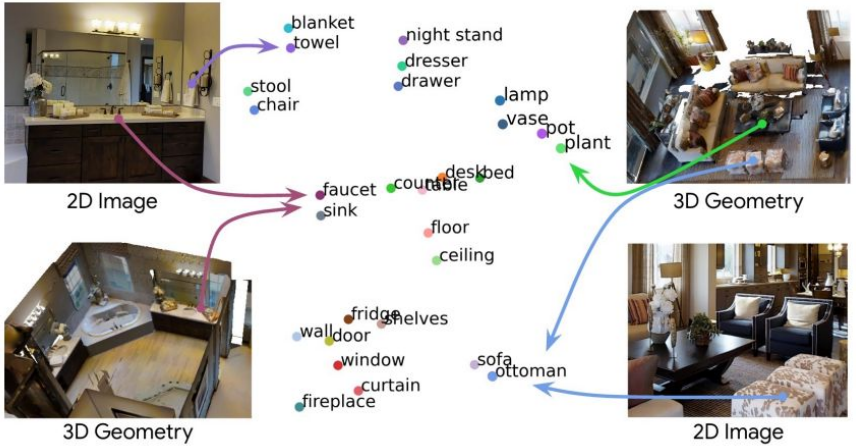
- VR/AR
- Robotics



2. Related work : OpenScene



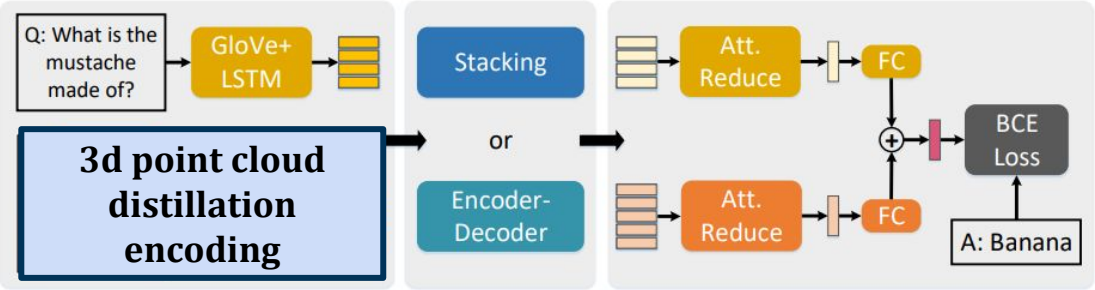
Method Overview



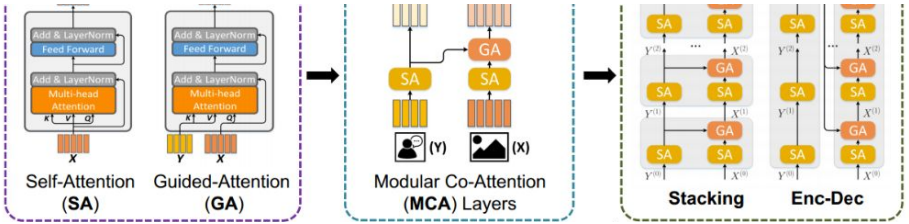
co-embed 3D points with text and image pixels in the CLIP feature space

[1] Peng, Songyou, et al. "Openscene: 3d scene understanding with open vocabularies." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

2. Related work: MCAN-VQA & CLIP-ViL



How about replace the 2d CLIP visual encoding with the 3d point cloud distillation encoding?



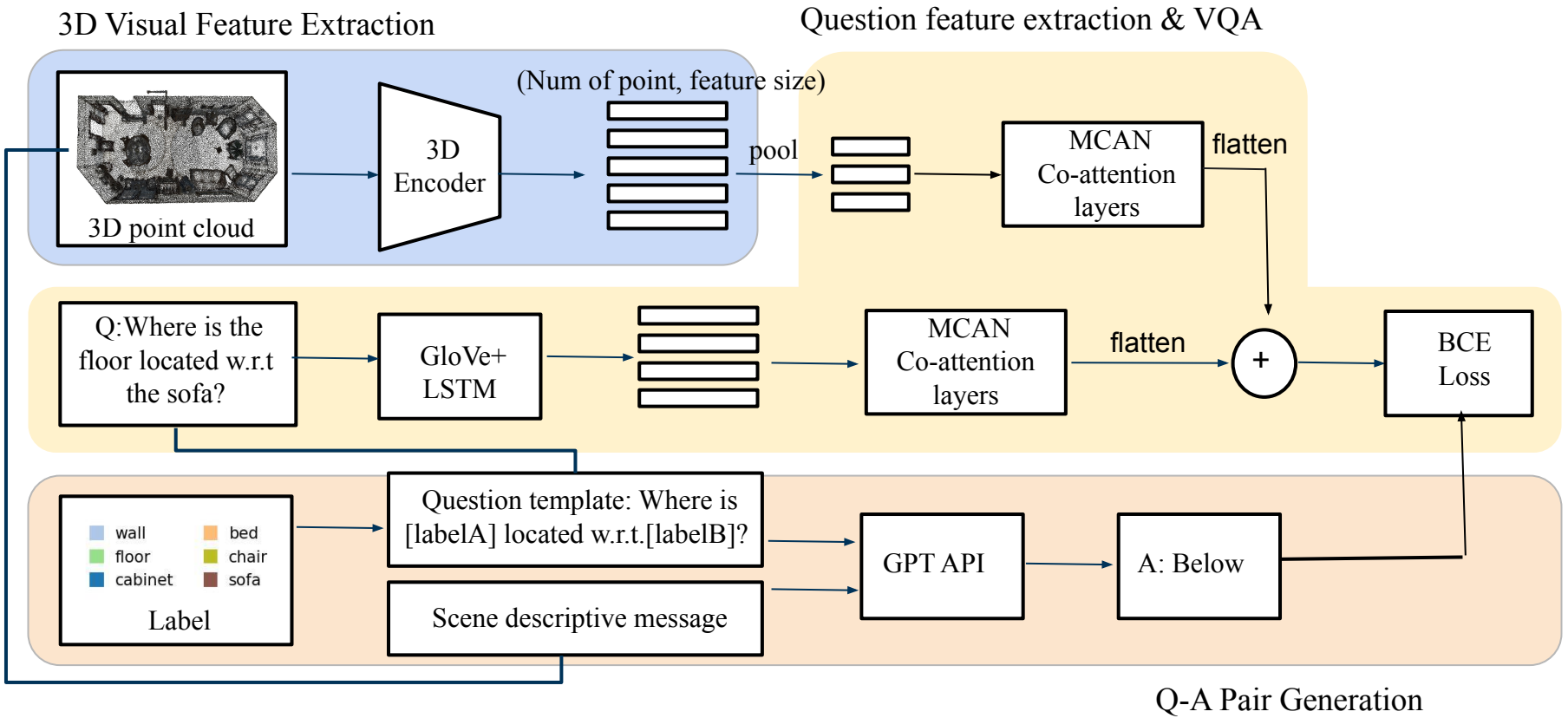
Stacked MCA layer

Pythia	ImageNet-Res50	61.54	61.76
	CLIP-ViT-B	51.38	51.47
	CLIP-Res50	65.55	65.78
	CLIP-Res101	66.76	67.14
	CLIP-Res50x4	**68.37**	**68.68**
MCAN	ImageNet-ResNet50	67.23	67.46
	CLIP-ViT-B	56.18	56.28
	CLIP-Res50	71.49	71.72
	CLIP-Res101	72.77	73.19
	CLIP-Res50x4	**74.01**	**74.17**

Result of CLIP-ViL

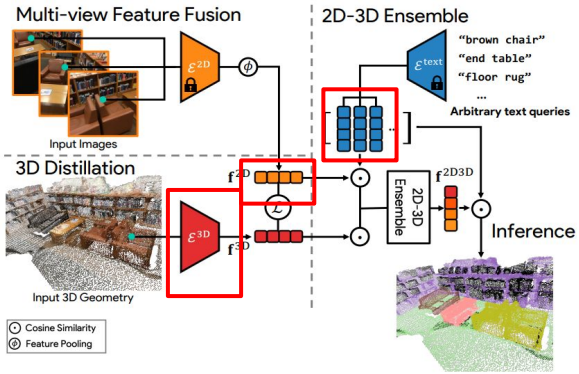
[2] Yu, Zhou, et al. "Deep modular co-attention networks for visual question answering." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[3] Shen, Sheng, et al. "How much can clip benefit vision-and-language tasks?." arXiv preprint arXiv:2107.06383 (2021)

3. Methodology: Overall framework

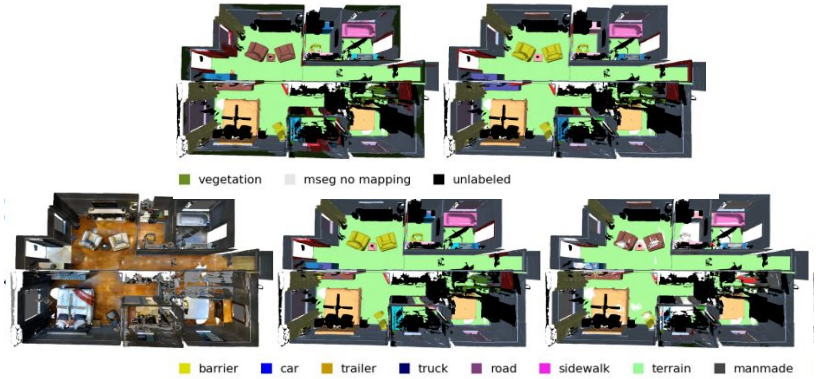


3. Methodology: 3D visual feature extraction

- Pretrained 3D distillation visual encoder: 3D feature (Matterport)
- Ablation study: Multi-/Singleview/2D-3D Ensembled feature
- Semantic segmentation results are used for pooling



Our idea: use the 3D distillation Visual Encoder



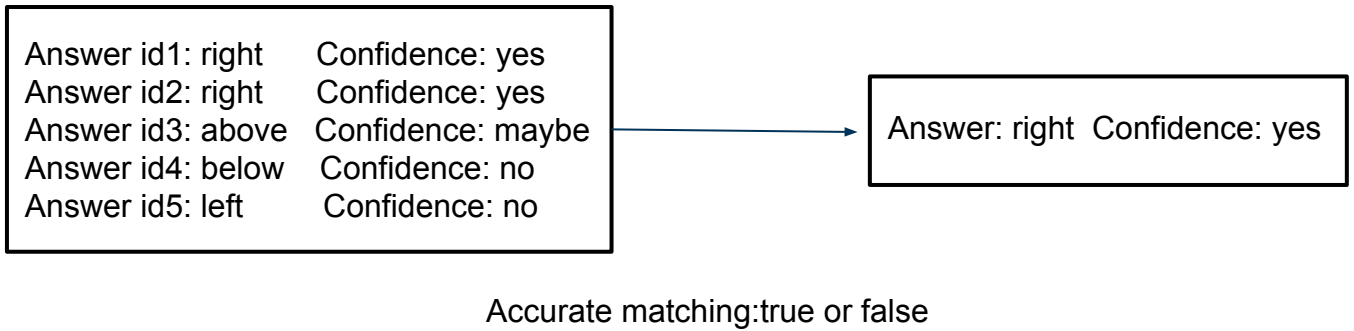
3D semantic segmentation results on public indoor benchmarks

[1] Peng, Songyou, et al. "Openscene: 3d scene understanding with open vocabularies." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[3] Chang, Angel, et al. "Matterport3d: Learning from rgb-d data in indoor environments." *arXiv preprint arXiv:1709.06158* (2017).

3. Methodology: Visual Question Answering

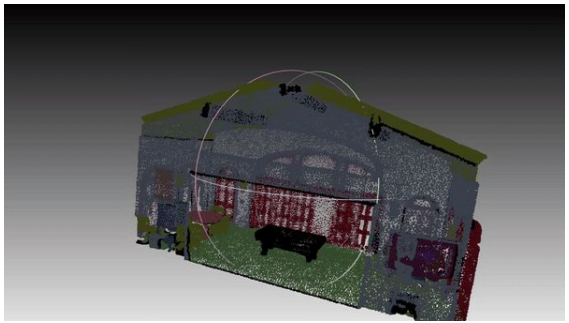
- Training: BCE loss
- Evaluation: per answer type accuracy
- From answer confidence to accurate matching



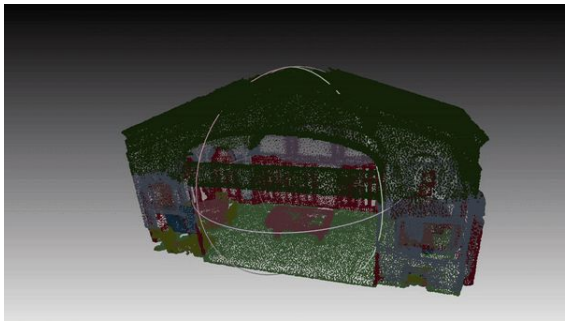
3. Methodology: Pooling Techniques

- 3D distilled feature size: (Num_Points, Feat_Dim), feature size for MCAN model: (Num_ROIs, Feat_Dim)
- Pooling techniques:
  1.Farthest Point Sampling + k-Nearest Neighborhood
  2.Semantic label based on ground truth
  3.Semantic label based on OpenScene prediction

FPS+kNN



GT semantic label

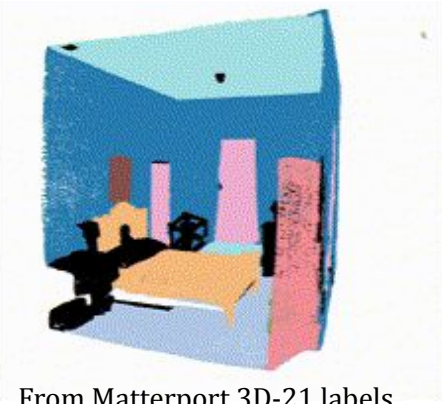


Predicted semantic label

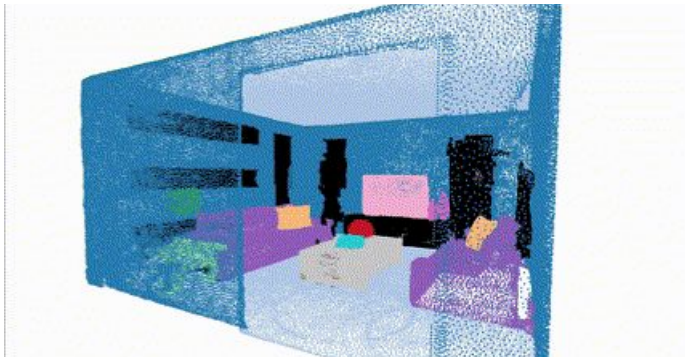
3. Methodology: Dataset Generation

- **MATTERPORT_LABELS_21** = ('wall', 'floor', 'cabinet', 'bed', 'chair', 'sofa', 'table', 'door', 'window', 'bookshelf', 'picture', 'counter', 'desk', 'curtain', 'refrigerator', 'shower curtain', 'toilet', 'sink', 'bathtub', 'other', 'ceiling')

- **MATTERPORT_LABELS_41** = ('wall', 'door', 'ceiling', 'floor', 'picture', 'window', 'chair', 'pillow', 'lamp', 'cabinet', 'curtain', 'table', 'plant', 'mirror', 'towel', 'sink', 'shelves', 'sofa', 'bed', 'night stand', 'toilet', 'column', 'banister', 'stairs', 'stool', 'vase', 'television', 'pot', 'desk', 'box', 'coffee table', 'counter', 'bench', 'garbage bin', 'fireplace', 'clothes', 'bathtub', 'book', 'air vent', 'faucet')



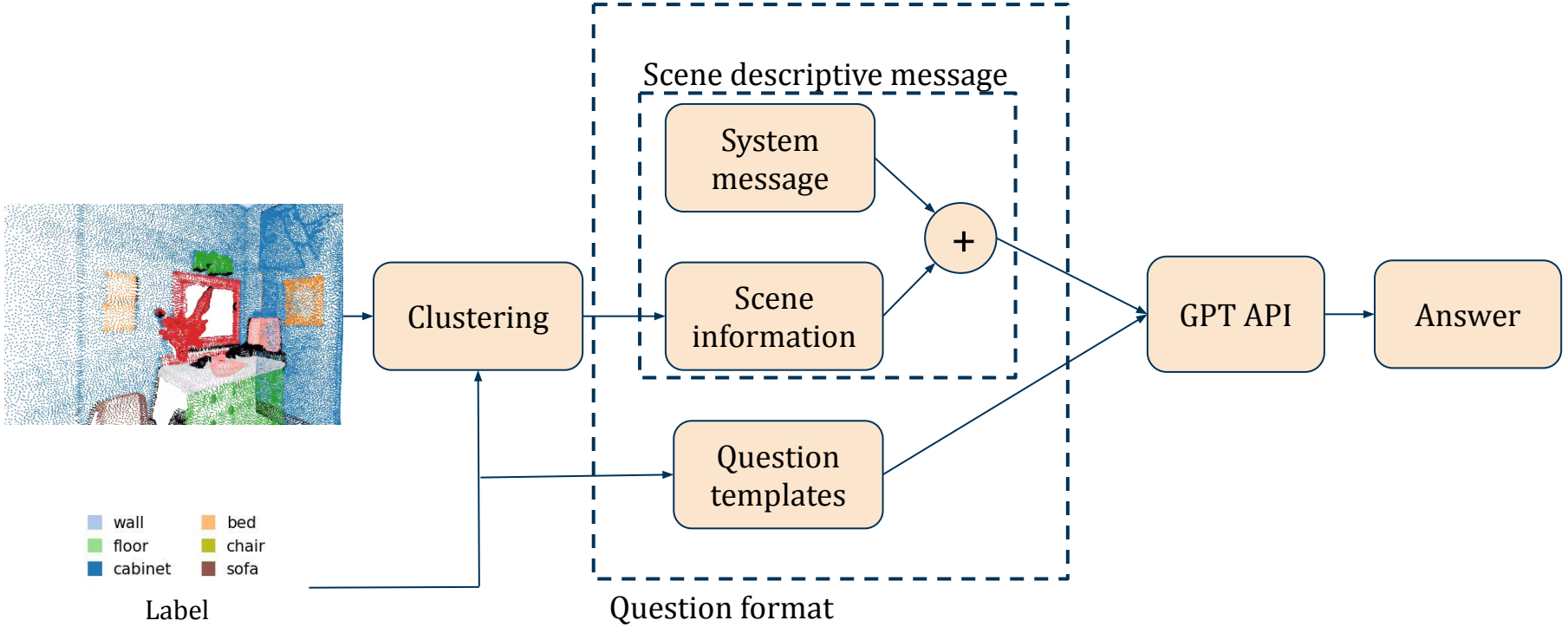
From Matterport 3D-21 labels



From Matterport 3D-41 labels

[3] Chang, Angel, et al. "Matterport3d: Learning from rgb-d data in indoor environments." *arXiv preprint arXiv:1709.06158* (2017).

3. Methodology: Dataset Generation



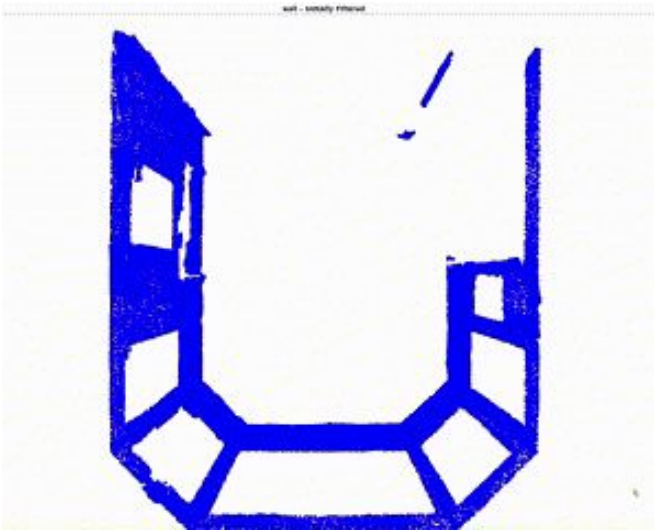
3. Methodology: Dataset Generation

- **For clustering the objects:**

Plane Segmentation

E.g. walls, floor, ceiling

Parameters are selected according to;
- Scan quality (sparse point cloud, noise in the point cloud etc.)



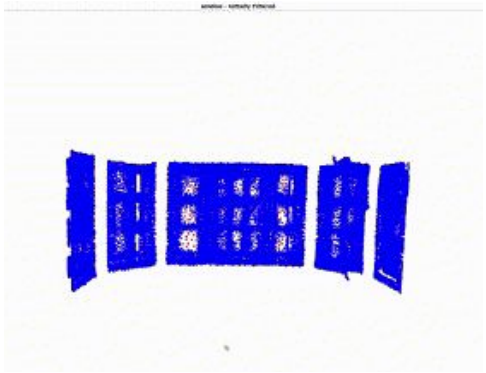
3. Methodology: Dataset Generation

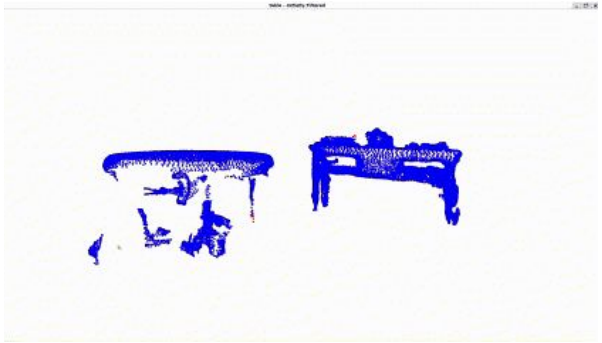
- **For clustering the objects:**

DB Scan

Other objects. E.g. table, picture, bed, vase, plant, lamp etc.

Parameters are selected according to;
- Object size
- Distance between objects
- Scan quality





3. Methodology: Dataset Generation

- **Question types:**

```
['count', 'spatial', 'yes/no', 'action']
```

- **Some question template examples:**

Spatial -> `"Where is the {obj1} located with respect to the {obj2}?`

Count -> `"How many {obj}s are there?"`

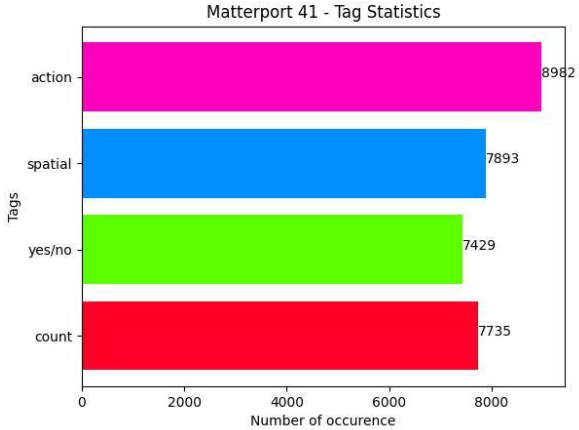
Yes/No -> `"Are there any {obj}s in the room?"`

Actions -> `"What is hanging on the wall?"`

4. Result: Our Dataset

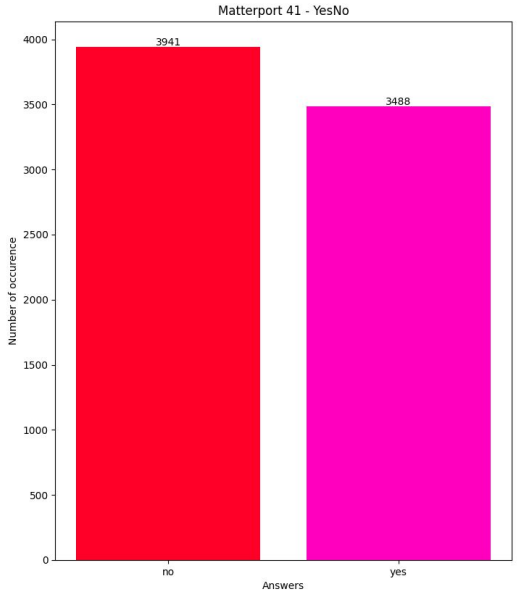
- **Training set:** 32039 Q&A pairs for 1025 different scenes
  - Yes/No: 7429
  - Count: 7735
  - Spatial: 7893
  - Action: 8982

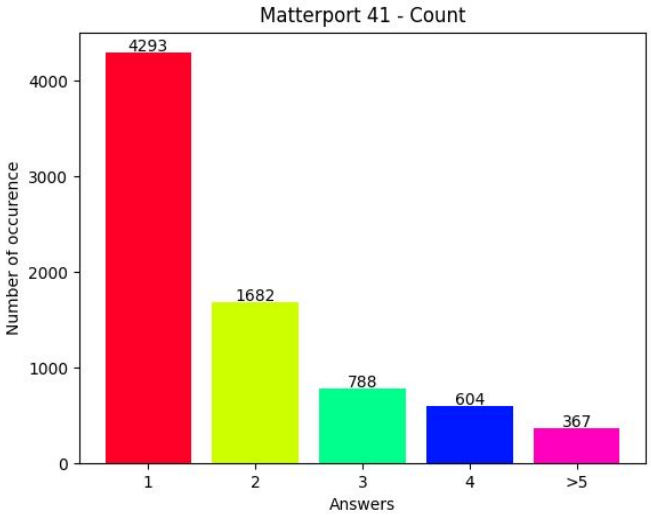
- **Validation set:** 5431 Q&A pairs for different 171 scenes.
  - Yes/No: 1249
  - Count: 1292
  - Spatial: 1364
  - Action: 1526

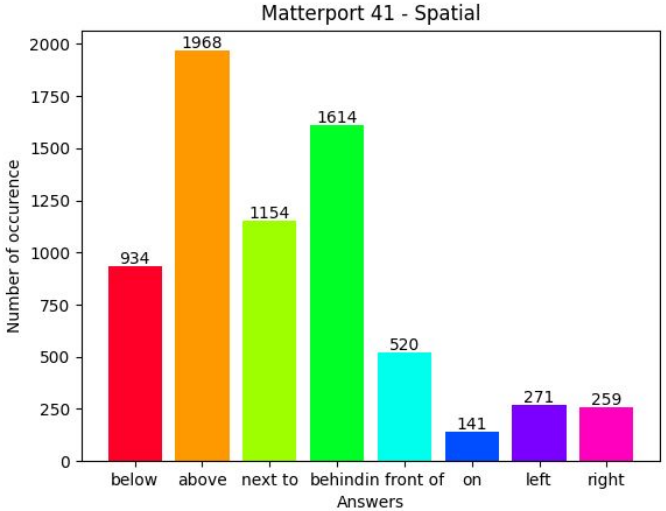


The Matterport 3D dataset is used for creating Q&A pairs.

4. Result: Our Dataset



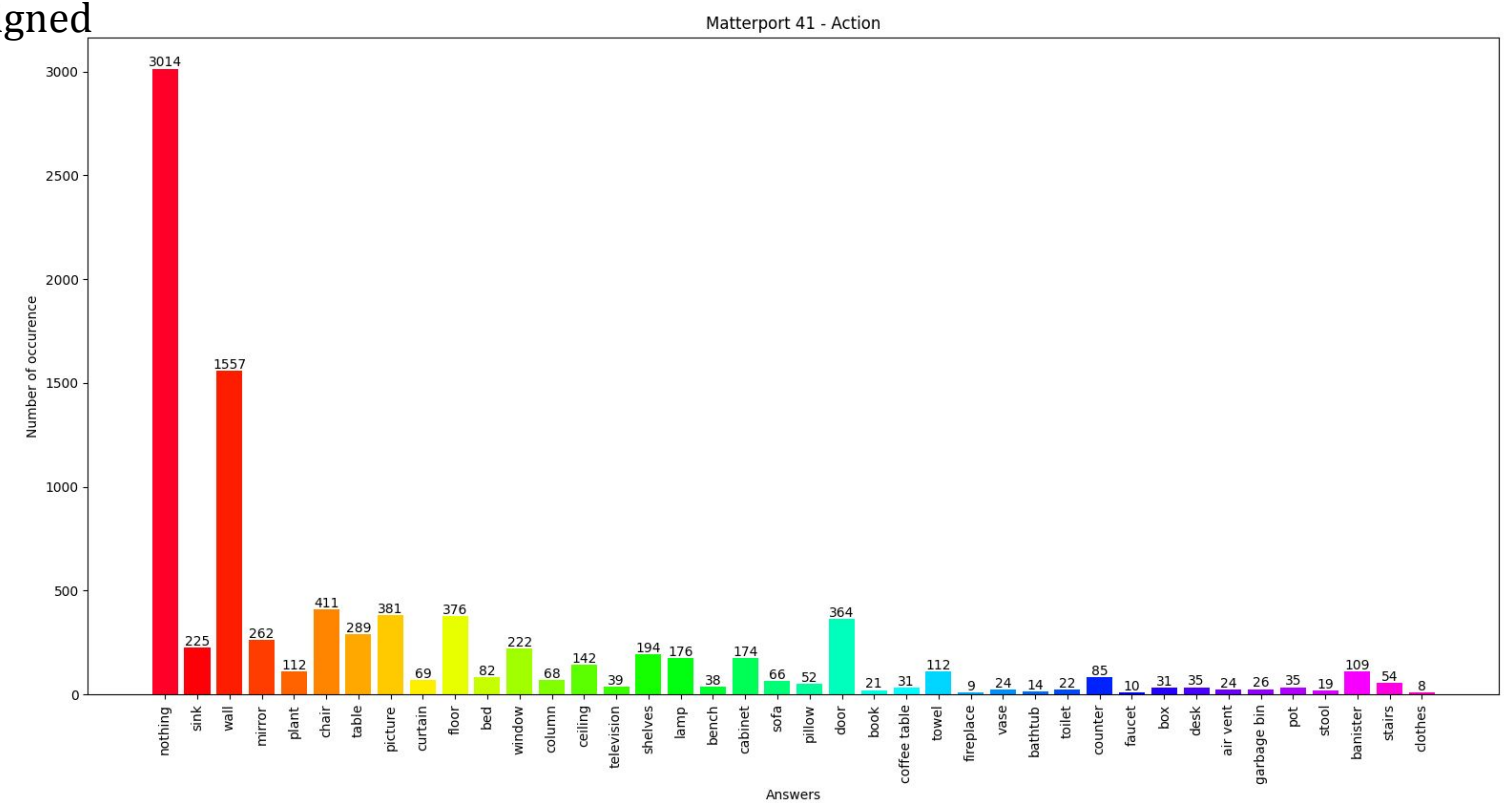




*Unaligned

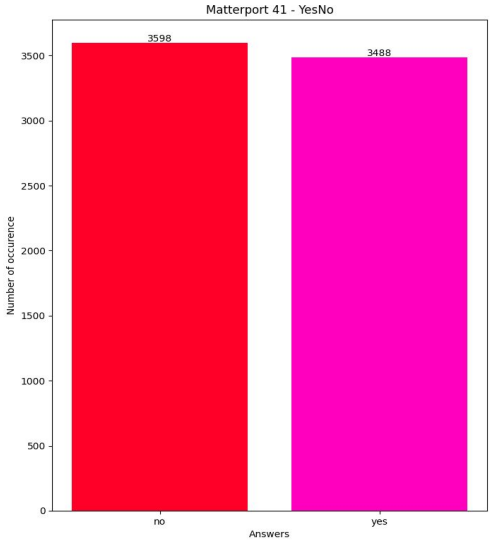
4. Result: Our Dataset

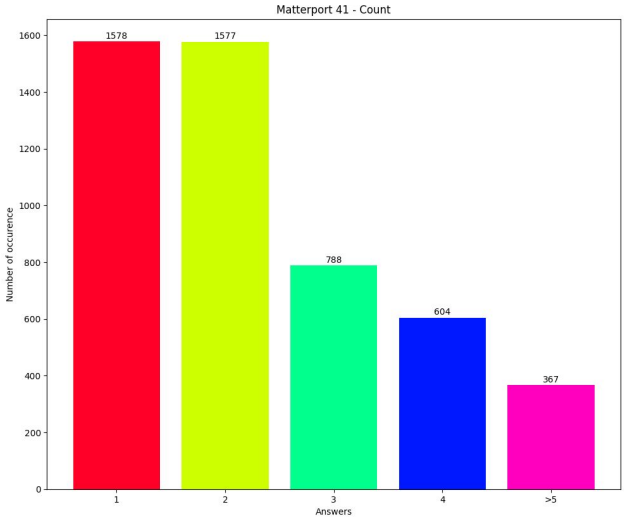
*Unaligned

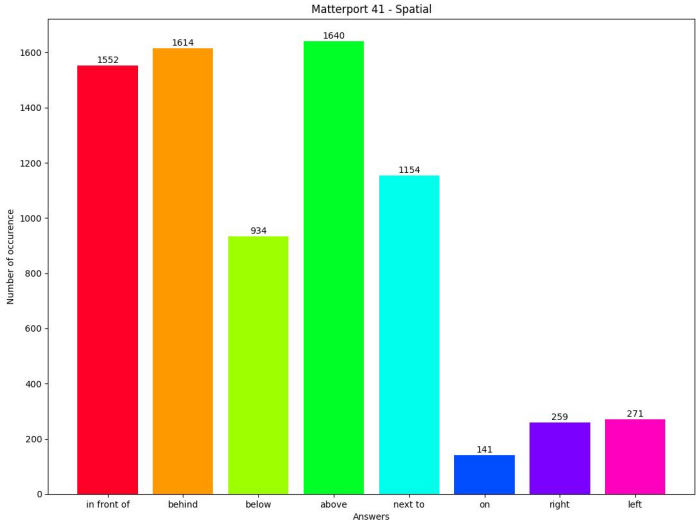


4. Result: Our Dataset

*Balanced version

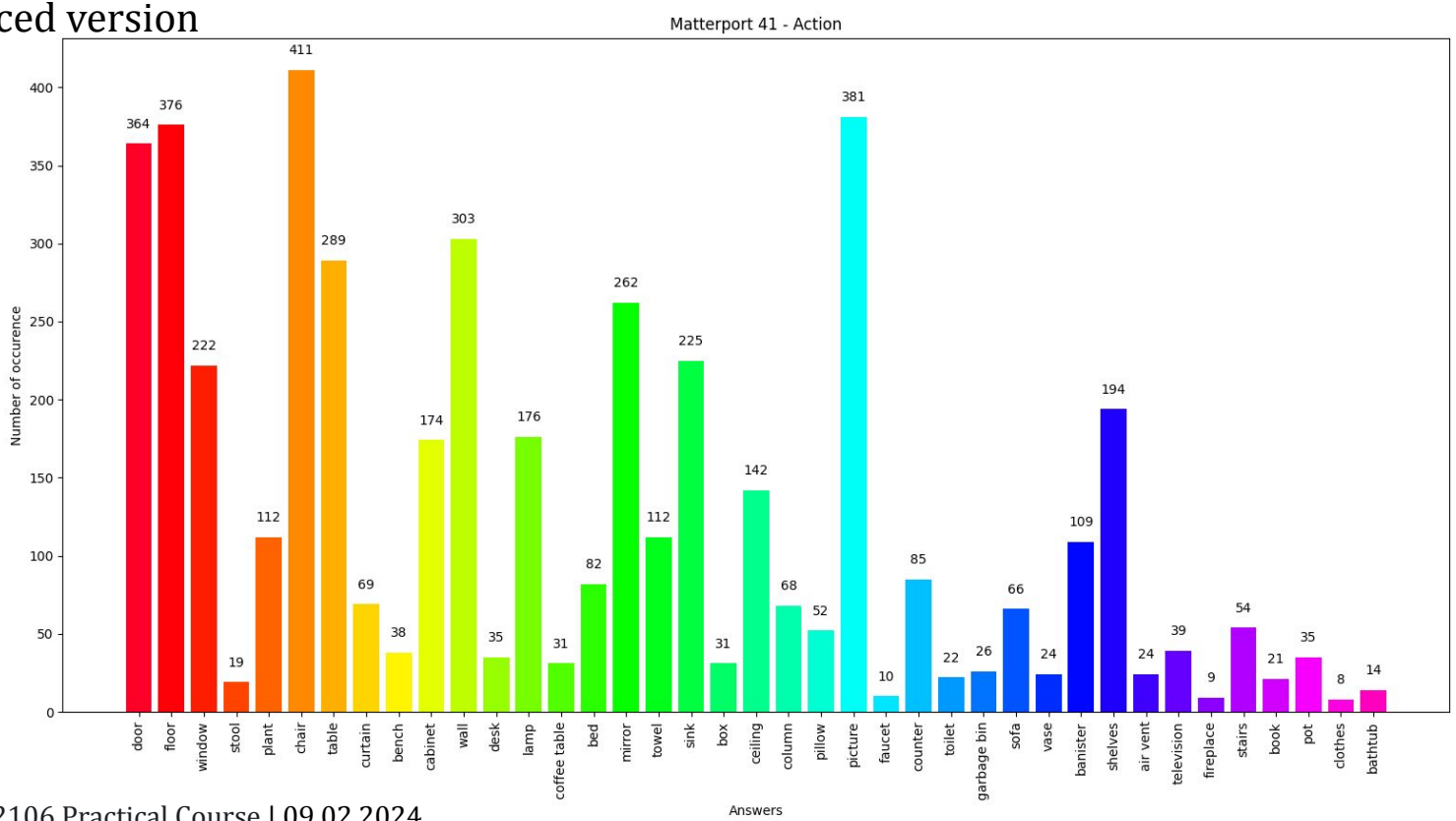






4. Result: Our Dataset

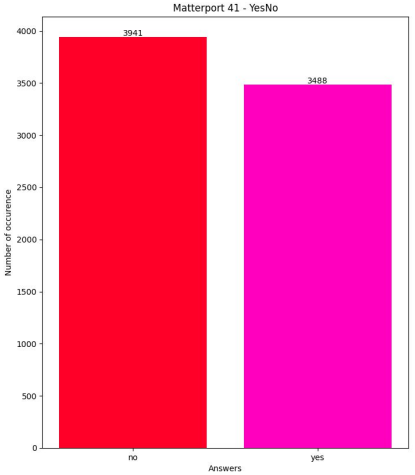
*Balanced version



4. Result: Demo

4. Result: Metric and Comment

- Metrics
  - Per-type Q&A accuracy and Overall accuracy
  - Higher is better



- Final result

Type	overall	count	spatial	yes/no	action
Accuracy(%)	58.73	56.9	48.02	93.58	41.30

- Does our method just memorize the high mode answer?

4. Result: Ablation

- Accuracy (%) Comparisons with other feature pooling methods

feature	dataset	pooling method	overall	count	spatial	yes/no	action
3D Distillation	Matterport 41 label	FPS	55.3	**57.88**	44.21	91.36	33.5
		GT semantic label	**58.82**	57.72	**48.02**	92.3	**41.99**
		predicted semantic label	58.73	56.9	48.02	**93.58**	41.3

4. Result: Ablation

- Accuracy (%) Comparisons with other feature types

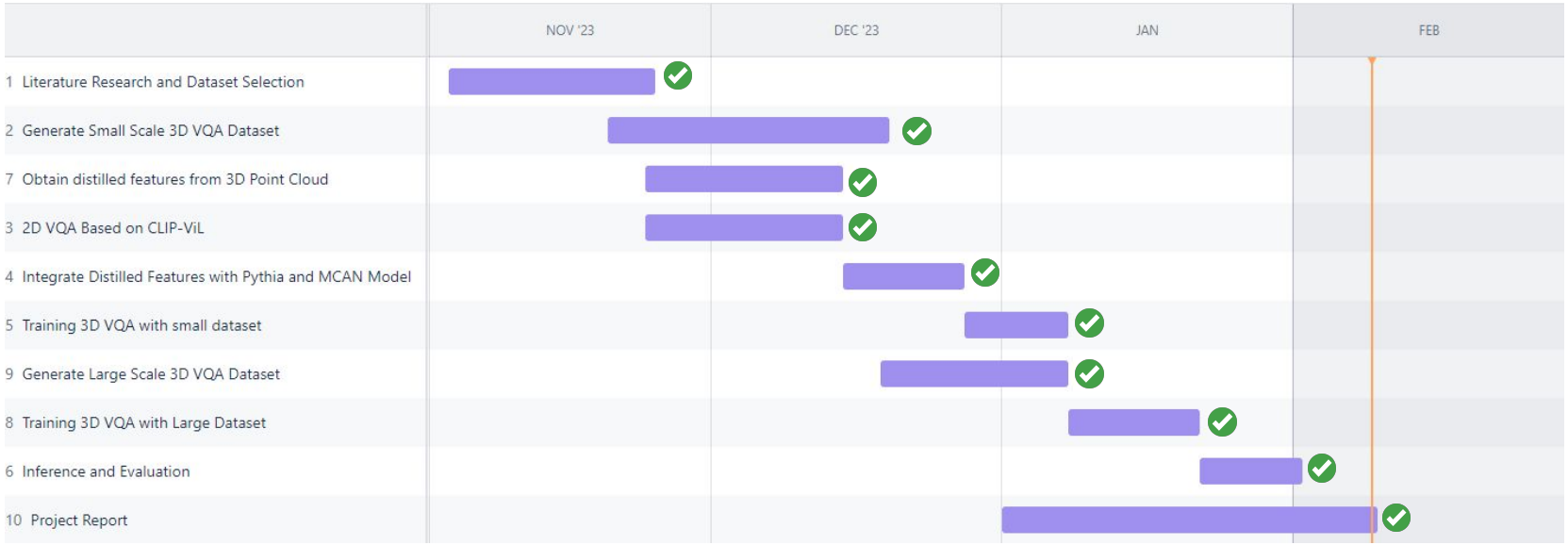
feature	pooling method	overall	count	spatial	yes/no	action
3D Distillation	GT semantic label	58.82	57.72	48.02	92.3	41.99
	predicted semantic label	58.73	56.9	48.02	93.58	41.3
2D Multiview	GT semantic label	59.08	58.24	50.79	94.14	40.32
	predicted semantic label	58.45	58.2	50.53	93.35	38.96
2D Singleview	GT semantic label	57.03	57.72	46.51	91.96	37.26
	predicted semantic label	55.85	52.38	44.76	92.3	38.86
2D-3D ensemble	GT semantic label	**61.7**	58.7	**52.54**	**94.78**	**45.33**
	predicted semantic label	61.14	**59.2**	52.14	93.76	44.15

5. Future Work:

- Using instance map to pool the feature
- Extend the model to different datasets
- Verify the dataset

6. Gains from the lecture:

- Theoretical knowledge of 3D vision
- Practical skills
- Team work



7. References

[1] Peng, Songyou, et al. "Openscene: 3d scene understanding with open vocabularies." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[2] Yu, Zhou, et al. "Deep modular co-attention networks for visual question answering." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[3] Shen, Sheng, et al. "How much can clip benefit vision-and-language tasks?." arXiv preprint arXiv:2107.06383 (2021).
[4] Qi, Charles Ruizhong tai, et al. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space." Advances in neural information processing systems 30 (2017).
[5] Zhu, Xiangyang, et al. "Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
[6] Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., & Gan, C. (2023). 3D-LLM: Injecting the 3D World into Large Language Models. *ArXiv*. /abs/2307.12981

Thank you!

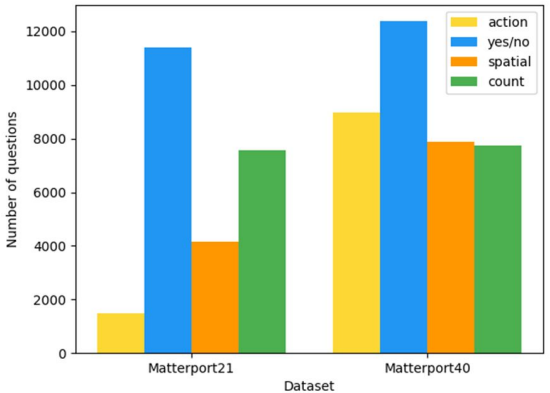
Questions?

Supplementary: Our Dataset

- **Training set:** 24581 Q&A pairs for 1467 different scenes
  - Yes/No: 11404
  - Count: 7558
  - Spatial: 4152
  - Action: 1467

- **Validation set:** 3744 Q&A pairs for different 225 scenes.
  - Yes/No: 1731
  - Count: 1157
  - Spatial: 631
  - Action: 225

- **MATTERPORT_LABELS_21** is used.



Supplementary: Results

- Comments on the Initial results

Type	count	spatial	yes/no	action	overall
Accuracy	62.4%	66.72%	93.66%	70.81%	77.99%

Q&A pairs with wrongly generated answers.
False positives.

Lack of Q&A pair variety, mostly have repeated questions.

The highest #Q&A pairs with correct generated answers.

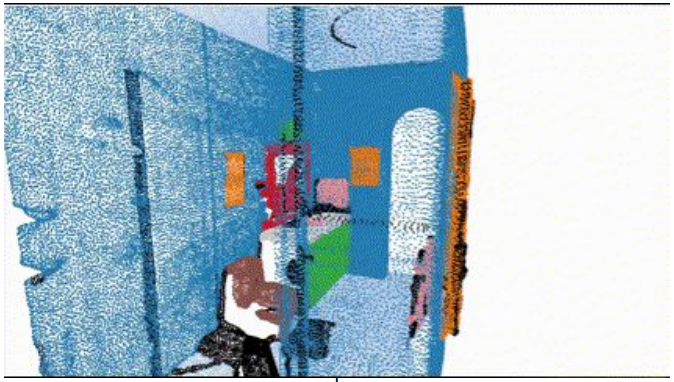
The lowest #Q&A pairs with correct answers.
Lack of variety.

- **MATTERPORT_LABELS_21** is used.

Supplementary: Results

feature	pooling method	# label	Overall	count	spatial	yes/no	action
3D distilled feature	fps	21	72.36	62.21	35.31	93.03	70.81
		41	59.89	47.7	44.13	93.05	38.86
	GT semantic label	21	77.68	62.21	66.89	93.03	70.81
		41	63.65	57.47	48.02	93.05	42.55
	predicted semantic label	21	77.65	62.21	66.72	93.03	70.81
		41	62.85	57.22	48.02	93.1	39.42
multiview	GT semantic label	21	79.11	63.57	65.08	94.95	77.25
		41	64.05	58.85	50.83	93.61	40.5
	predicted semantic label	21	76.56	59.1	66.04	93.03	70.81
		41	61.44	53.12	47.14	93.05	37.95
single view	GT semantic label	21	72.36	62.21	35.31	93.03	70.81
		41	62.92	57.72	47.3	93.05	39.97
	predicted semantic label	21	70.65	56.72	35.31	93.03	70.81
		41	61.71	54.35	46.03	93.05	39

Supplementary: Question Generation



Clustering

```
Scene_id,Label,Xc,Yc,Zc,Xmin,Ymin,Zmin,Xmax,Ymax,Zmax,Dx,Dy,Dz
oLBMNvg9in8_region5,door,0.780,0.530,-1.168,0.671,-0.063,-2.343,0.918,1.000,-0.220,0.246,1.063,2.122
oLBMNvg9in8_region5,door,1.600,-0.955,-1.210,0.926,-1.565,-2.332,2.019,0.017,-0.163,1.093,1.582,2.169
oLBMNvg9in8_region5,door,-0.171,1.276,-1.243,-0.932,0.683,-2.335,0.183,1.933,-0.142,1.116,1.249,2.193
oLBMNvg9in8_region5,door,-1.135,2.404,-1.267,-1.440,1.850,-2.326,-0.422,2.888,-0.362,1.018,1.037,1.964
oLBMNvg9in8_region5,wall,-1.173,-0.255,-1.219,-1.287,-1.082,-2.326,-0.930,2.872,-0.131,0.357,3.954,2.195
oLBMNvg9in8_region5,wall,-0.078,-1.057,-0.928,-1.022,-1.106,-2.326,0.946,-1.007,-0.183,1.968,0.099,2.143
oLBMNvg9in8_region5,wall,1.233,-0.036,-1.097,-0.989,-0.139,-2.234,1.877,0.022,-0.040,2.866,0.161,2.193
oLBMNvg9in8_region5,wall,-0.223,0.937,-1.381,-1.480,0.878,-2.341,0.743,0.991,-0.212,2.223,0.113,2.129
oLBMNvg9in8_region5,floor,-0.005,0.392,-2.309,-1.461,-1.557,-2.349,1.959,2.891,-2.203,3.421,4.449,0.146
oLBMNvg9in8_region5,floor,0.341,0.034,-0.088,-0.432,-0.981,-0.128,1.507,1.682,-0.019,1.939,2.663,0.109
oLBMNvg9in8_region5,ceiling,-0.087,0.386,-0.196,-1.555,-1.233,-0.278,1.911,2.609,-0.135,3.466,3.842,0.143
oLBMNvg9in8_region5,window,0.042,2.219,-0.714,-0.036,1.900,-1.271,0.145,2.480,-0.447,0.181,0.580,0.824
oLBMNvg9in8_region5,window,0.006,2.222,-1.175,-0.045,2.058,-1.324,0.042,2.421,-1.049,0.087,0.363,0.276
oLBMNvg9in8_region5,book,-1.295,1.202,-1.479,-1.560,0.926,-1.795,-1.187,1.395,-1.268,0.373,0.469,0.527
oLBMNvg9in8_region5,stairs,-0.702,0.319,-0.073,-0.750,-0.138,-0.193,-0.680,0.770,-0.007,0.070,0.908,0.186
oLBMNvg9in8_region5,stairs,-0.830,0.316,-2.191,-0.951,-0.096,-2.317,-0.789,0.758,-2.096,0.162,0.854,0.221
```

1. Scene information

Supplementary: Question Generation

Algorithm Question Generation

Inputs:

1. *Scene ID, Label, X, Y, Z (Scene info)*
2. *System Message*

for *i* **in** *num_of_questions:*

1. *Select question type*
2. *Select labels (objects)*
3. *Create question template*
4. *Question format = scene info+system message+ question template*

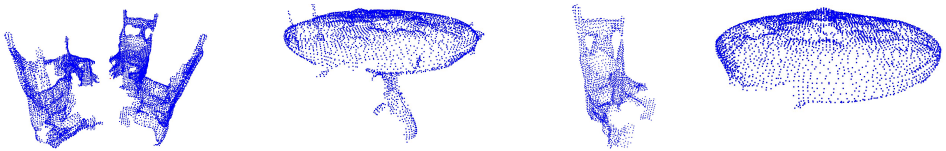
end

Output:

Question format

Supplementary: Challenges

- Hardware bottleneck
- Finding the best DbScan parameters



`eps=0.25, min_samples=4` `eps=0.1, min_samples=2`

- Incorrect point clouds