ИНТЕЛЛЕКТ AIRI FusionBrain.Robotics

Lecture 5

3D models: MLLMs application for 3D CV

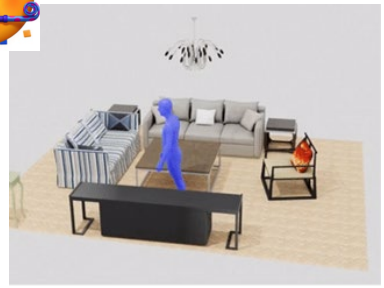
Zinkovich Viktoriia

Œ Ś Ľ ½ ↕ ½ ╤ ╗ È Ë ╤ Ë Ñ ╤ Ĵ È ↕ ╤
! Ĉ ↕ Ś Ë Ś È Ë ½ ╤ Ë Ś Ľ Ĵ Ĉ Ś Ë ╤ Ś ╤

Introduction: Course Plan

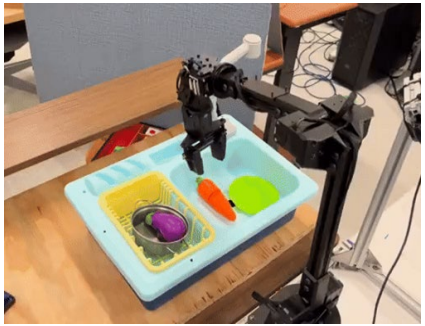
Day 1
Image modality

Day 3
Data generation
in MLLMs



Day 5
3D models

Day 2
Video modality

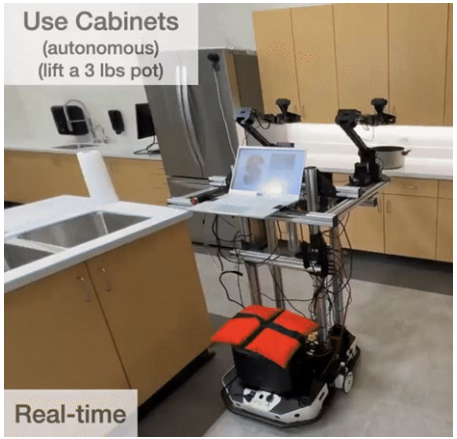


Day 4
Action modality

Recap: Lecture #4

many embodiments characterized by **how we can control it**

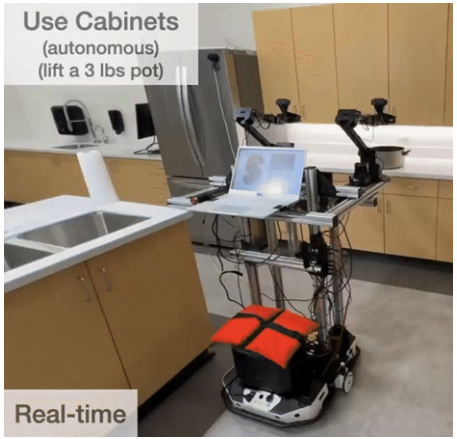
Recap: Lecture #4



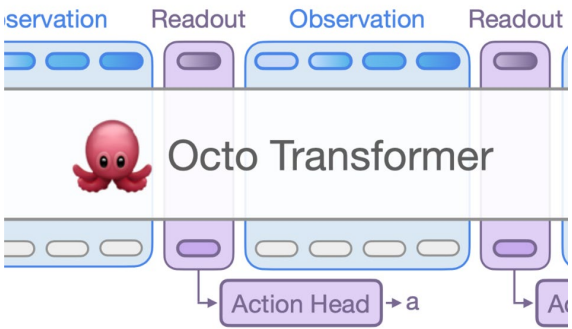
evaluation can be done by **humans** , teleoperation, data-gloves

Recap: Lecture #4

many embodiments characterized by **how we can control it**



evaluation can be done by **humans** , teleoperation, data-gloves



policies to train models for robotic tasks: RT-1, RT-2, Octo, Diffusion, OpenVLA

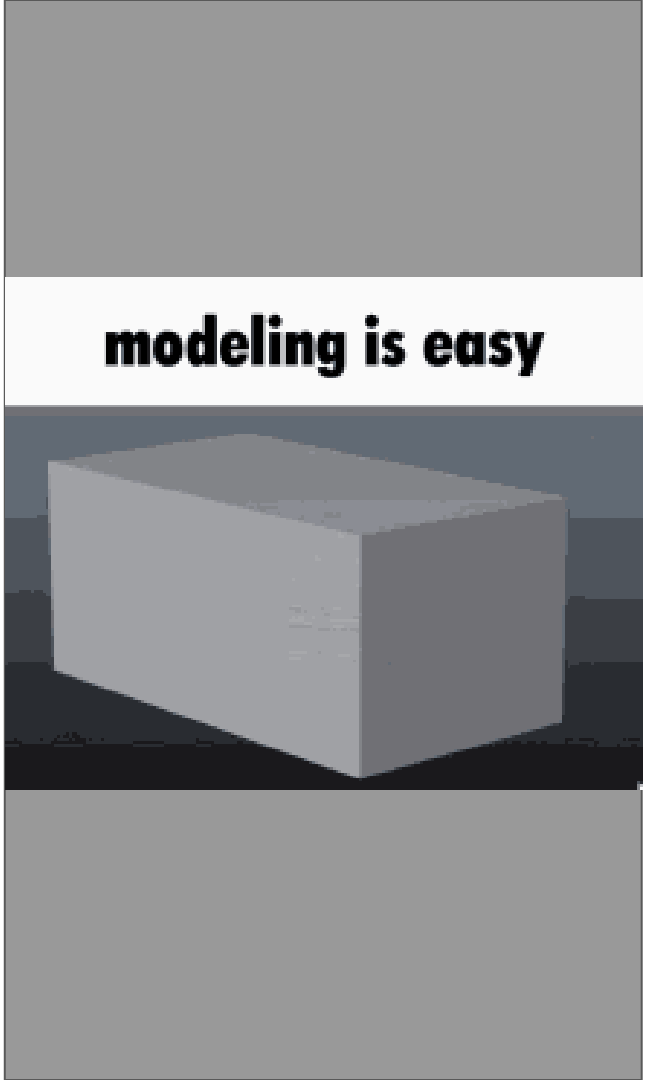
ÜŚĽĨĴ ĈŚË ╚½╤

1 What is **3D vision** ?

2 How we get **information** about 3D?

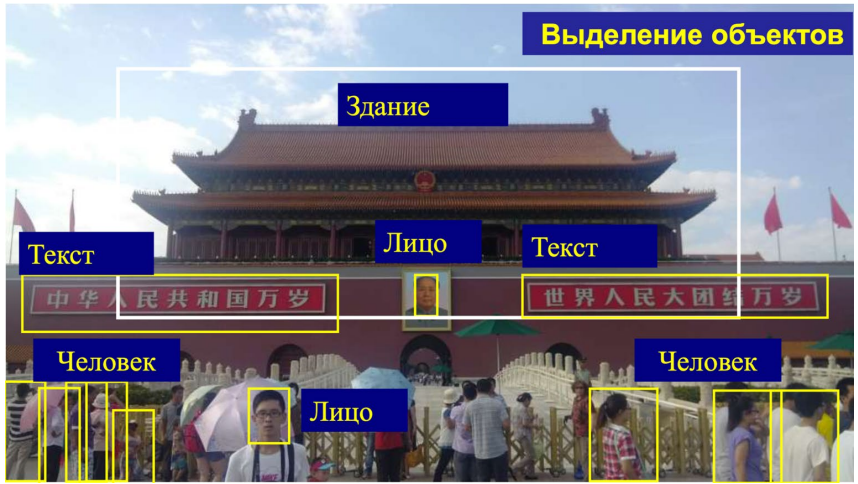
3 Tasks & **Datasets**

4 ° ĹĖĴ ½╚ÜÜê Ė



Introduction: What is 3D vision?

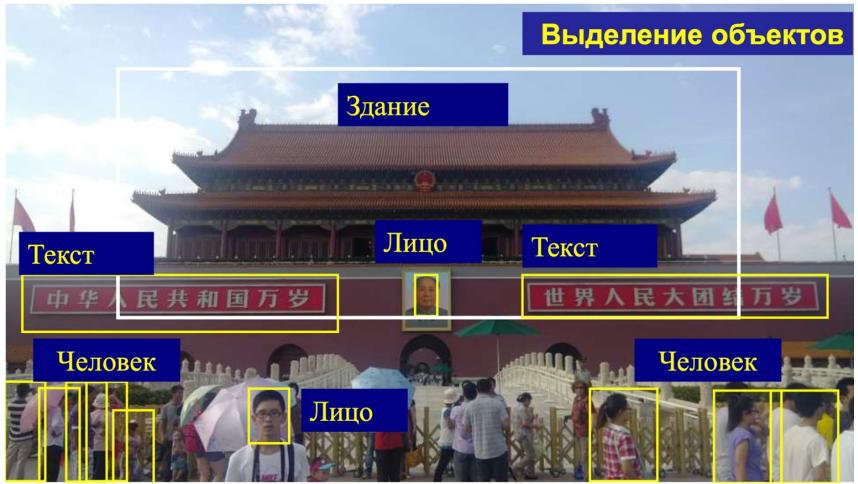
The Turing test (1949 originally) for **computer vision** — answer any question about the image that a human can answer



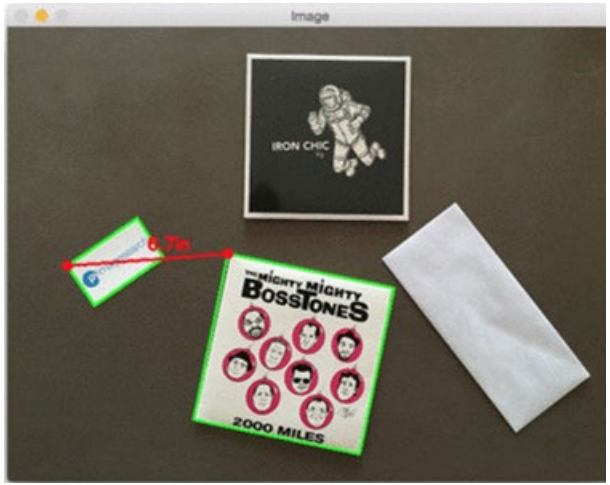
semantic questions
(where / how many / text)

Introduction: What is 3D vision?

The Turing test (1949 originally) for **computer vision** — answer any question about the image that a human can answer



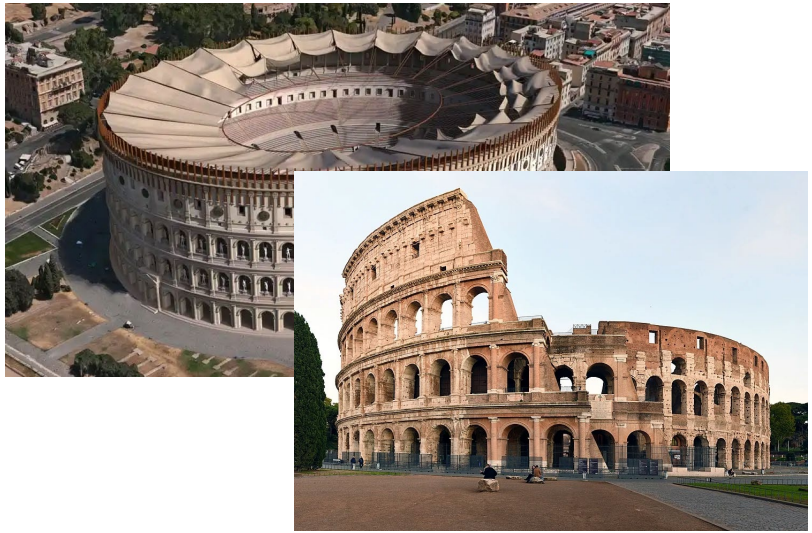
semantic questions
(where / how many / text)



metric questions
(shape / distance)

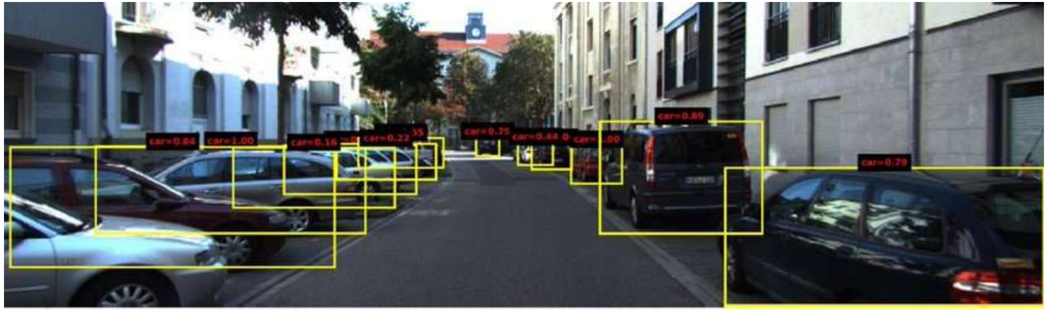
Introduction: What is 3D vision?

The difficult task is **total 3D reconstruction** , we need to build an accurate 3D model that allows: 1) manipulation with objects; 2) image synthesis

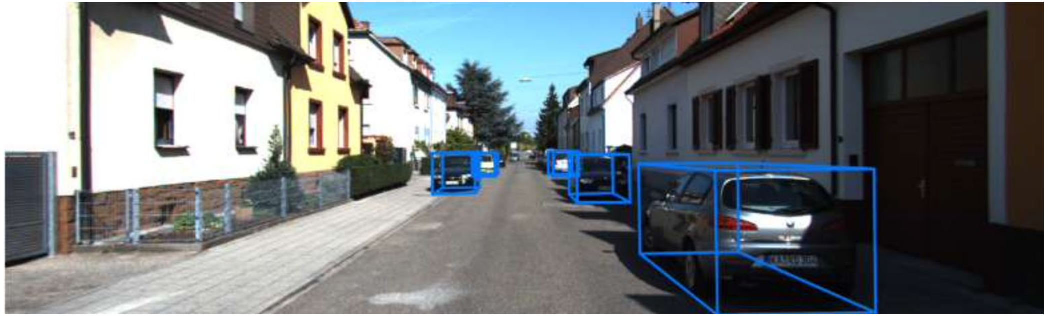


p╤ĨǬ ň Ĵ ĽĨ ╤ÏË ↕½ĨËĖËAB Ëu ĖË ╤Ḱ

2D



3D



3D vision =

Metric Vision + Semantic Vision

1

k▌ŬH▌Ë ŚĨË▬!▌Ĉ▌ ½Ĩ▌▬Ë ½Í▌Ĵ ĨËABḰ

Sensors, LiDARs…

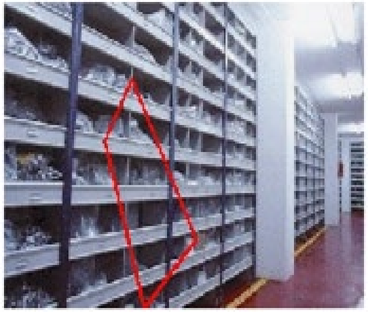
3D Info: 1. Structure from Motion

Find key points on the images, define camera motion — **"Structure from Motion"** → sparse point cloud that can be further improved (SIFT + RANSAC)

Agarwal et. al. Building Rome in a Day. ICCV 2009

3D Info: 1. Structure from Motion

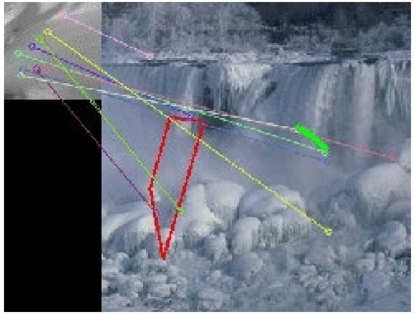
Find key points on the images, define camera motion — **“Structure from Motion”** $\rightarrow$ sparse point cloud that can be further improved (SIFT + RANSAC)



Ground truth

SIFT+RANSAC

Ground truth



SIFT+RANSAC

Agarwal et. al. Building Rome in a Day. ICCV 2009

3D Info: 1. Structure from Motion

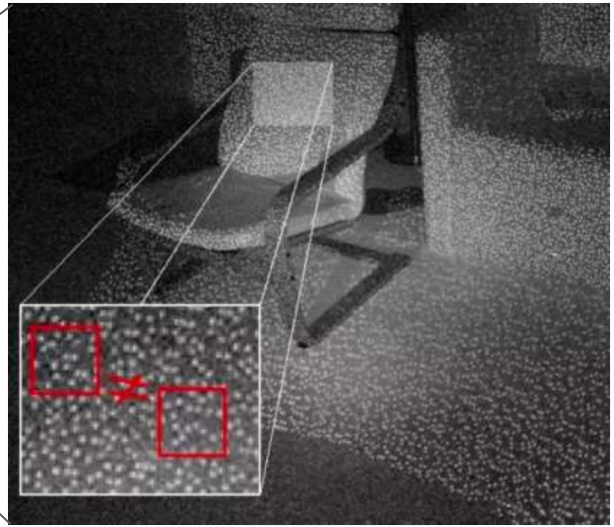
The problem with sparse reconstruction: **sparse** and **inaccurate** (noisy data) + **distortions** (the walls)

ÁB Ëþ╤!▌ ÏË SËB Ś◦ Ĩ↕Ë, ½╣ ŚĈ⁄₂Ė

RGB image + map of distances to objects → depth of the point pattern distortion

PrimeSense technology, licensed by **Microsoft** and implemented in the Kinect for **Xbox 360** camera

Projector



"Smart" **structural illumination** — set of spots according to a tricky pattern

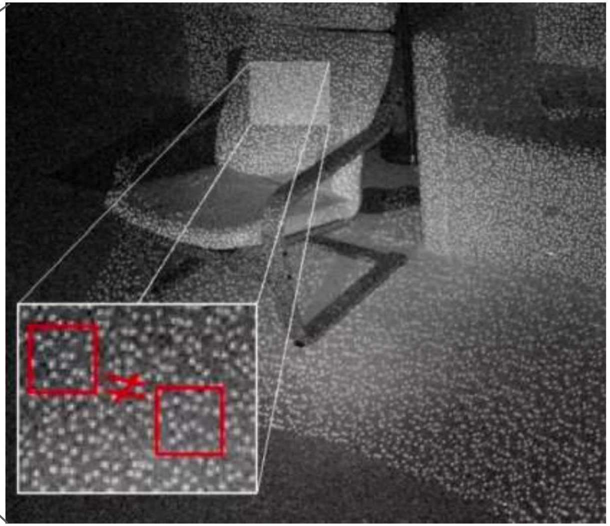
3D Info: 2. Depth Cameras

RGB image + map of distances to objects → depth of the point pattern distortion

PrimeSense technology, licensed by **Microsoft** and implemented in the Kinect for **Xbox 360** camera

Projector

Time-of-Fight
(Kinect-2)



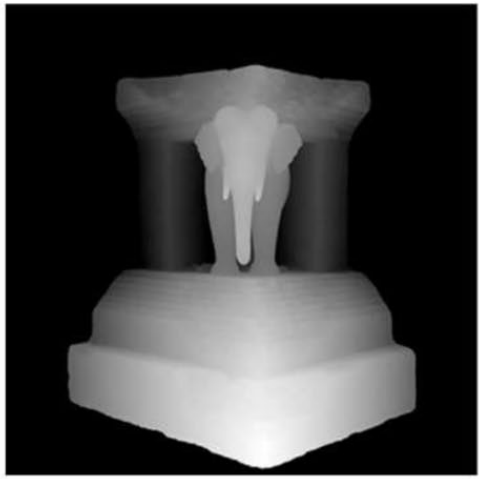
"Smart" **structural illumination** — set of spots according to a tricky pattern

3D Info: 2. Depth Cameras

→ RGB image is combined with depth (D) to **obtain RGBD — CNN net!**

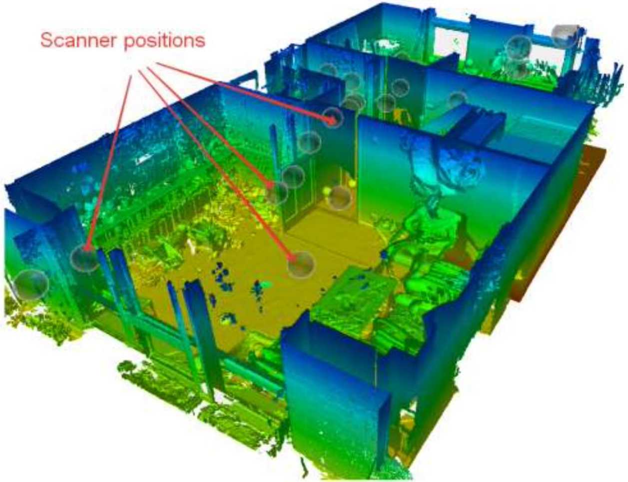
→ To combine depth with RGB need to **calibrate** camera poses

→ Sensors have **threshold** for depth estimation



ΆB Ëp╤!▌ ÏËASEÜ½ĖŚĈËŒĽ'½╤╤ŚĈĖ

Unlike RGB or depth cameras, **LiDARs** immediately provide a full 3D representation, not just a depth map — **time difference** between the emission and reception of the signal (phase shift)

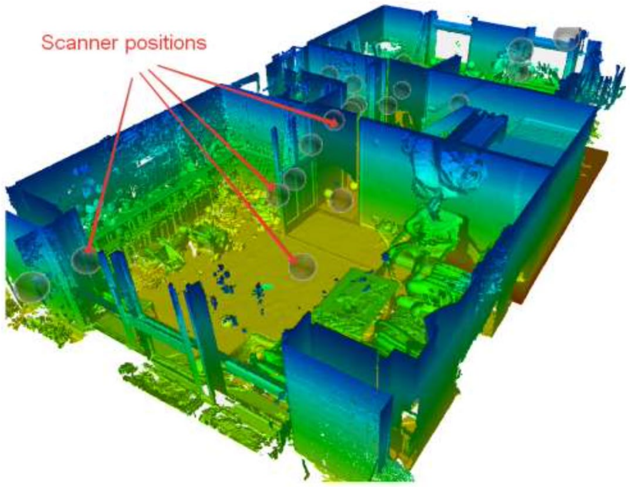


3D Info: 3. Laser Scanners

→ [illegible]

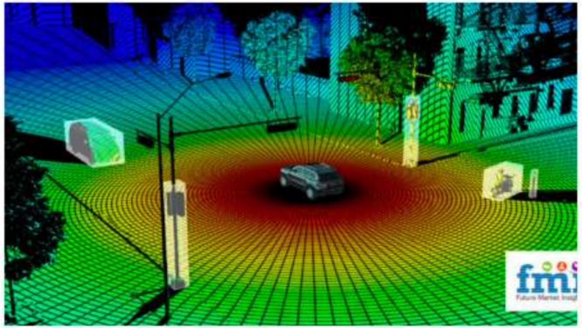
→ **Sparsity of data**
(where the laser did not reach the point)

→ **Solid -state LiDARs**
(autonomic vehicles)



3D Info: 4. Point Cloud

- [illegible]
- [illegible]
- [illegible]



2

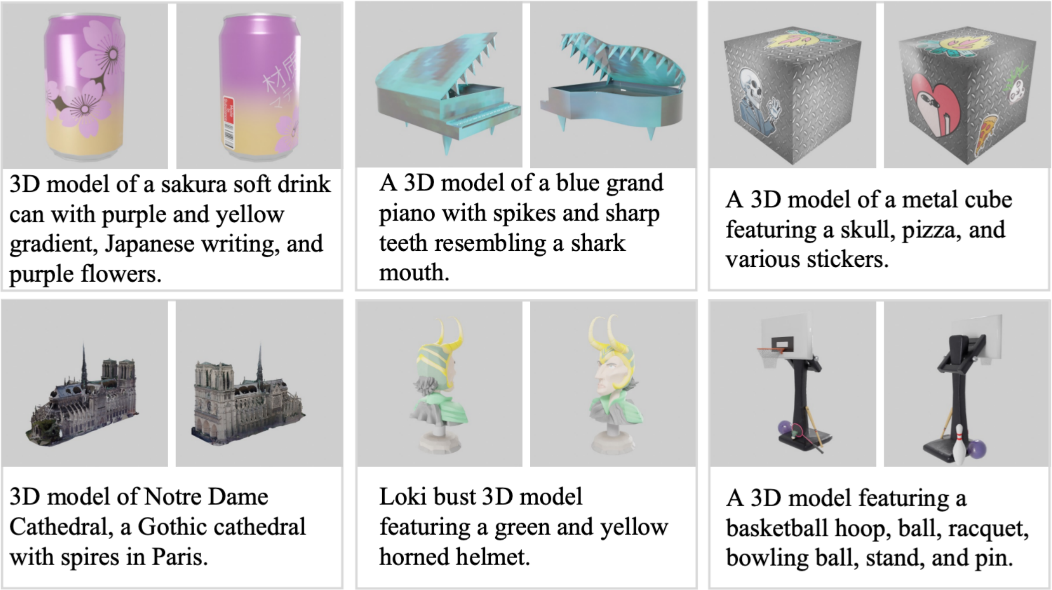
3D tasks and datasets

Captioning, segmentation

Tasks & Datasets: 3D Captioning

[illegible]

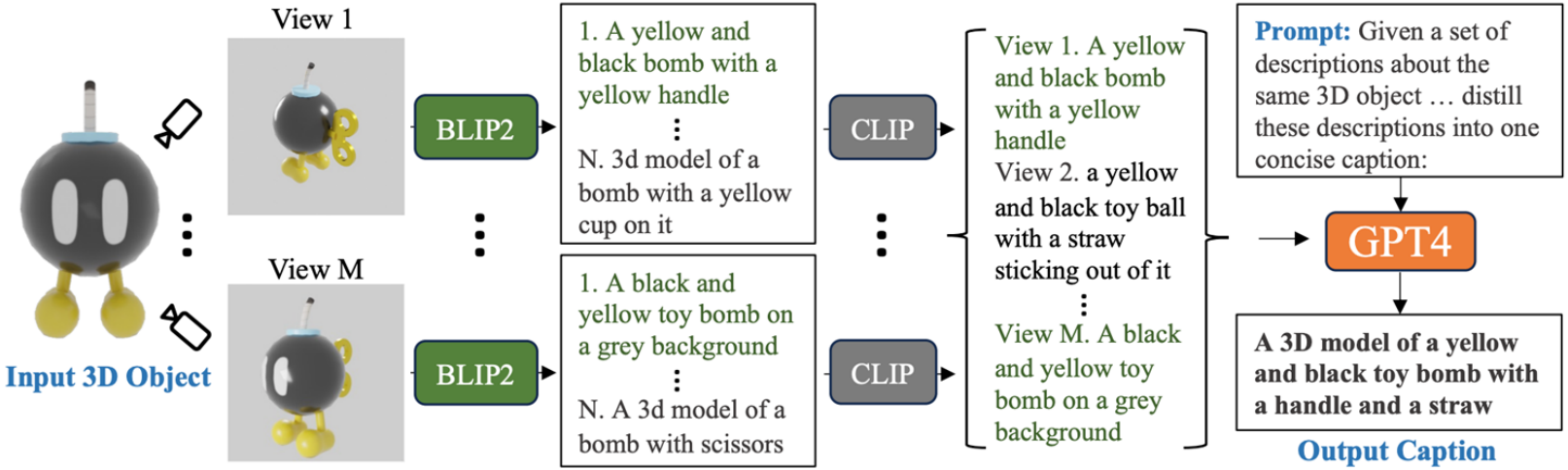
- **Cap3D dataset** (1.5 M) (based on ObjaVerse for generation of 3D objects)
- Cap3D needs **8 views**
- **Synthetic** captioning with VLM



Tiange Luo et al. Scalable 3D Captioning with Pretrained Models. NeurIPS 2023.

Tasks & Datasets: 3D Captioning

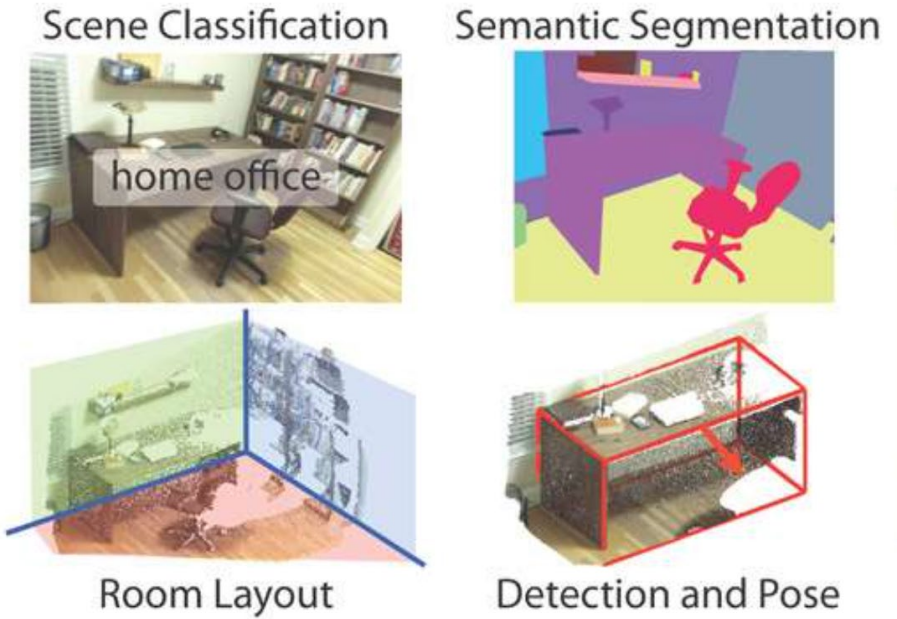
Because all components (BLIP2, CLIP, GPT-4) are pretrained, Cap3D requires **no human annotation** and scales linearly with the number of 3D assets



Tiange Luo et al. Scalable 3D Captioning with Pretrained Models. NeurIPS 2023.

Tasks & Datasets: SUN RGB-D (2015)

- **3D variant** of detection and segmentation tasks
- **10,000 images** from different cameras
- each scene was **shot 1 time**, without video
- **58,657** bounding boxes



S. Song, S. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite, CVPR 2015

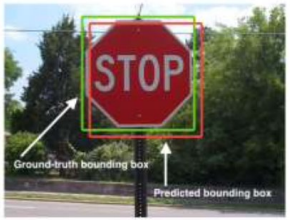
Tasks & Datasets: SUN RGB-D (2015)

- [illegible]
- [illegible]

Total Scene Understanding

S. Song, S. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite, CVPR 2015

Tasks & Datasets: SUN RGB-D (2015)

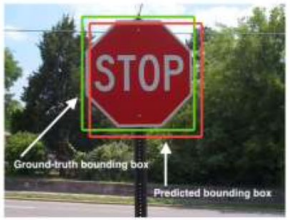


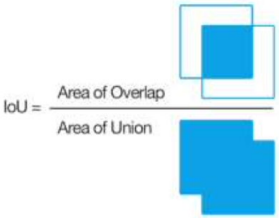
$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

How to calculate IoU metric in 3D ?

S. Song, S. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite, CVPR 2015

Tasks & Datasets: SUN RGB-D (2015)





Oriented IoU — matching in Z-axis + intersection of 2 projected polygons

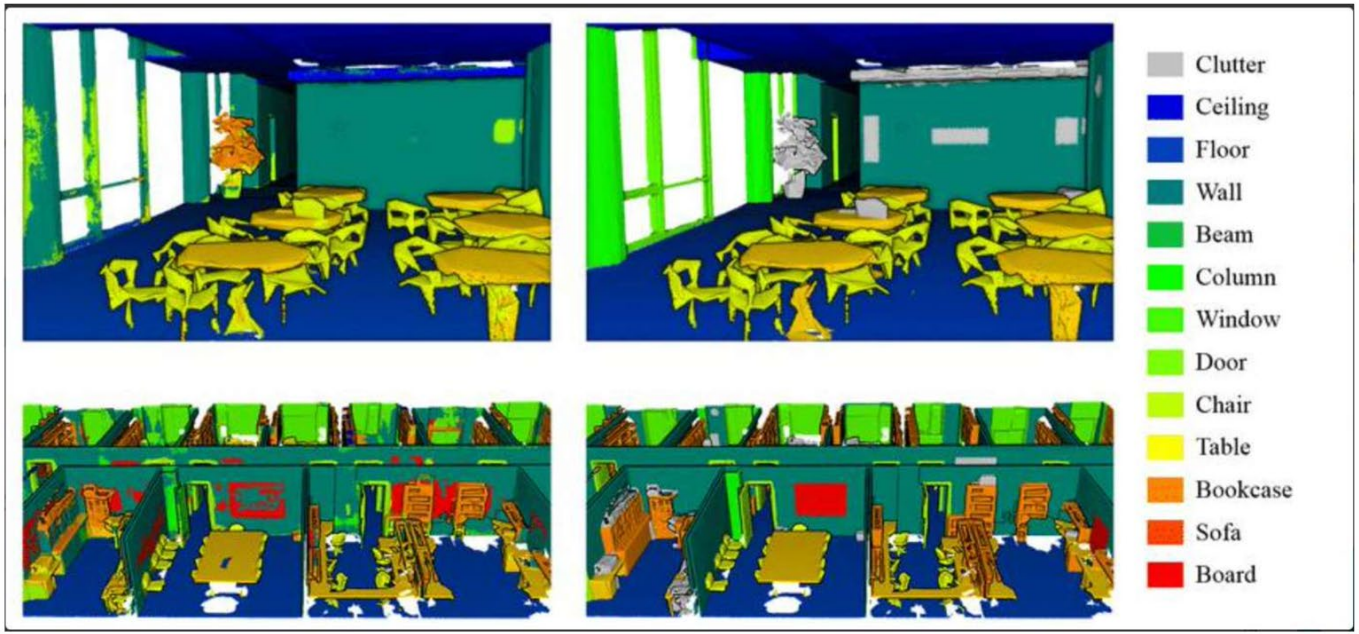
Average Precision (AP) — for all classes of objects and several IoU thresholds, mAP@0.25 and mAP@0.50

S. Song, S. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite, CVPR 2015

Tasks & Datasets: 3D Segmentation

3D Segmentation:

classification of scene points – you must put a class label



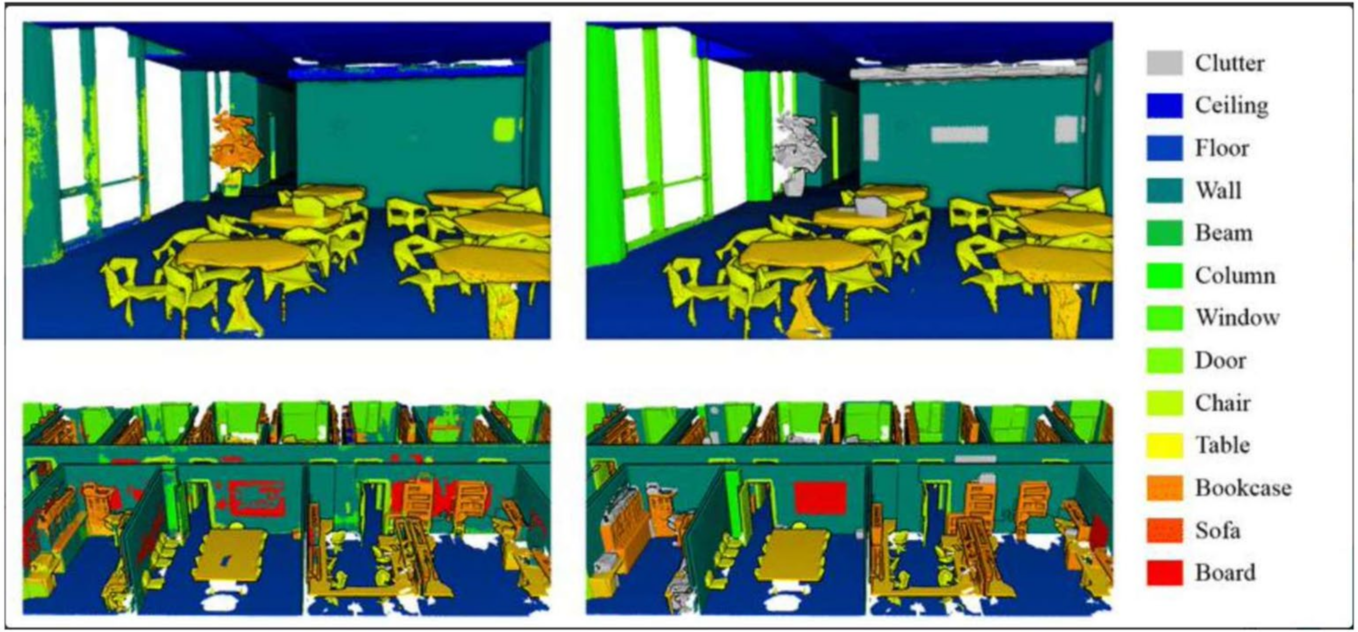
S. Song, S. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite, CVPR 2015

€½Ė╥Ė ┌ ЁB ½Ĩ ½ĖŚĨ ĖĨĖAB ĖŒŚn ╢ Ś╤Ĩ½Ĩ ╡ ╤

3D Segmentation:

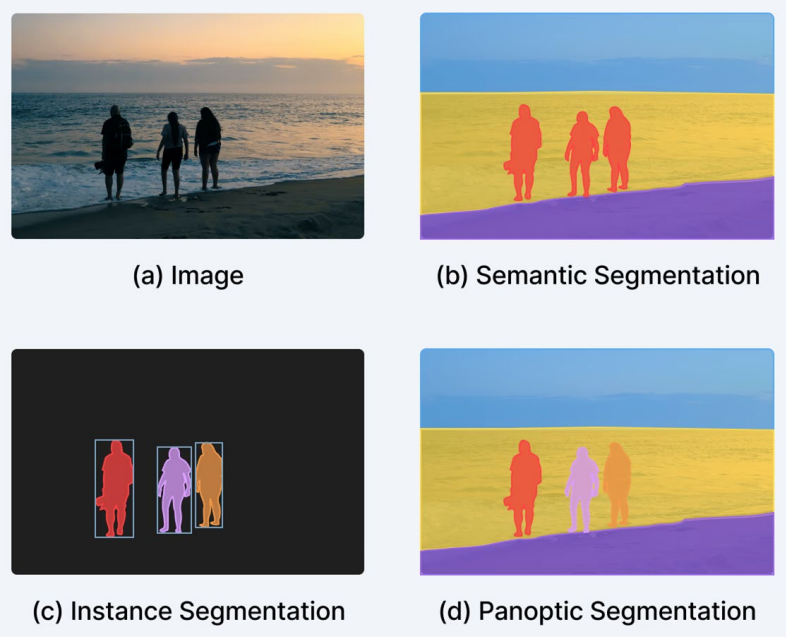
classification of scene points – you must put a class label

Instance Vs Panoptic?



S. Song, S. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite, CVPR 2015

Tasks & Datasets: Instance, Panoptic



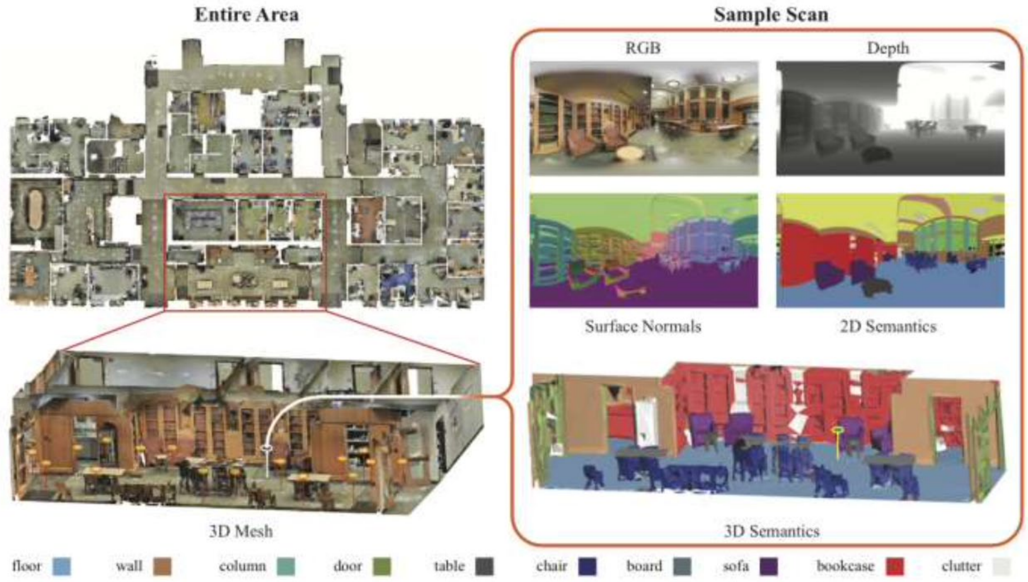
- **Instance** — cloud markup points with labels of individual instances of objects
- **Panoptic** — instance + semantic segmentations

Tasks & Datasets: S3DIS (2016)

- Scanning **6 large - scale areas** on the Stanford University campus
- Total of ≈**272 rooms** — offices, classrooms, corridors, conference rooms
- **215 million three - dimensional** XYZ coordinates with RGB color

Armeni et.al. 3D Semantic Parsing of Large-Scale Indoor Spaces, CVPR 2016.

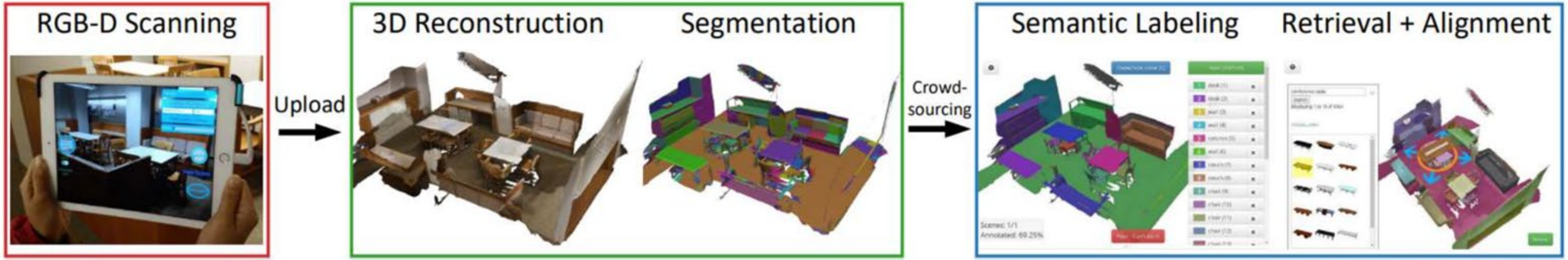
Tasks & Datasets: S3DIS (2016)



- **13 semantic classes** (wall, floor, ceiling, column, etc.)
- **Instance tags:** each individual object is assigned a unique ID
- **Markup files** are supplied for each combined cloud.txt with XYZRGB and two-label markup

Armeni et.al. 3D Semantic Parsing of Large-Scale Indoor Spaces, CVPR 2016.

Tasks & Datasets: ScanNet (2017)

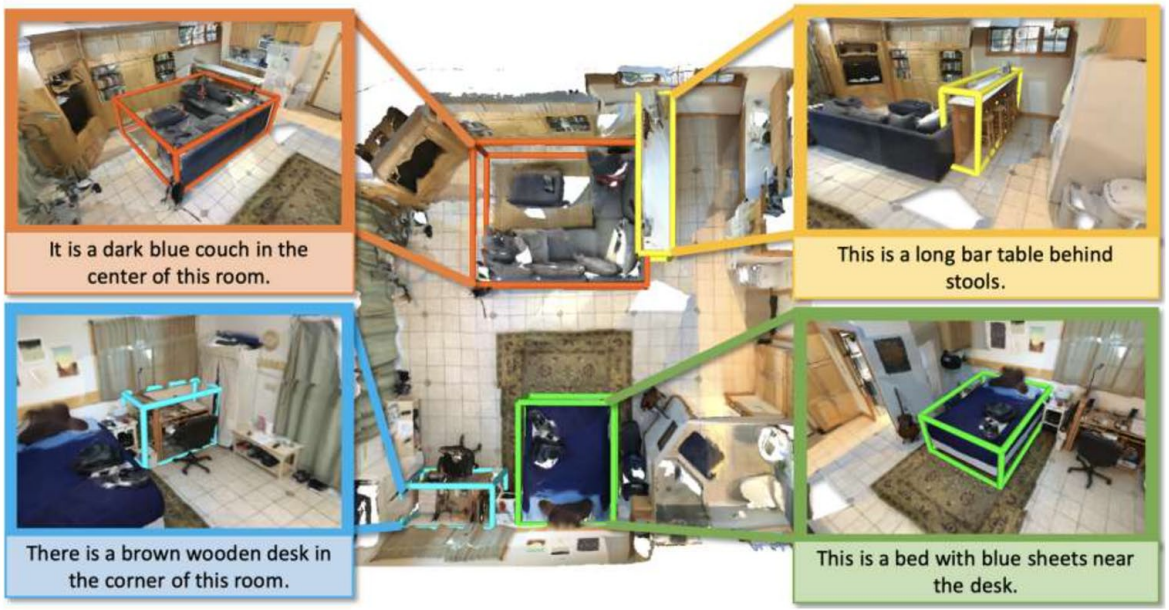


- [illegible]
- [illegible]

Tasks & Datasets: 3D Visual Grounding

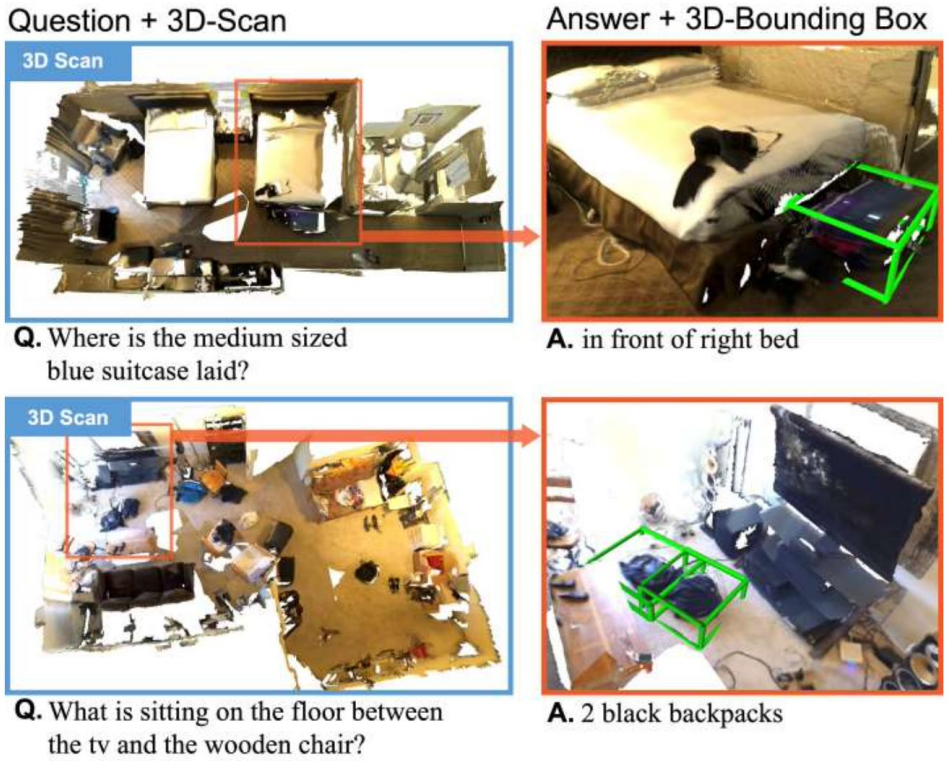
It's difficult to mark up rooms again, we can make **add –ons on top of this dataset** — localization of an object by text description

Tasks & Datasets: ScanRefer (2020)



- **51,583 text descriptions** for 11,046 objects
- **800 ScanNet** scenes
- Manual markup via **Amazon Turk**
- Checking by **students** – they need to find the object
- Still one of the **main datasets** for 3D visual grounding

Tasks & Datasets: 3D Question Answering



- **Extension** of 3D visual grounding
- Properties of objects, their **relative location**, and how to **find** group of objects
- **Answer** can be **different** (not only BB): 1) text + BB, 2) text, 3) text + multiple BB

Tasks & Datasets: ScanQA (2022)

Split	# Question	# Unique Question	# 3D Scenes
Train	25,563	20,546	562
Val	4,675	4,306	71
Test w/ objects	4,976	4,552	70
Test w/o objects	6,149	5,484	97
Total	41,363	32,337	800

Table 2. ScanQA dataset statistics.

- ScanNet + ScanRefer + additional markup
- incorrect / irrelevant questions

Underspecified questions

Q: What is in the corner?
- Several objects at corners!

Q: What color is the chair?
- Three chairs at the scene!

Valid questions

Q: What is over the chair beneath the blackboard?
- Answer: jacket

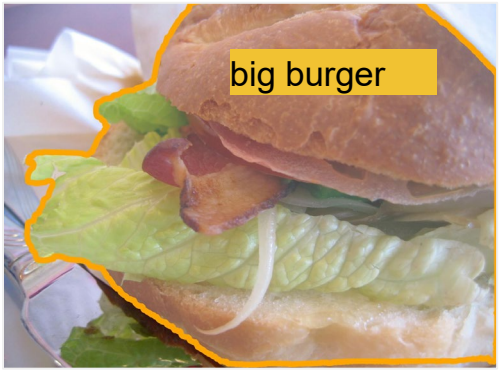
Q: What color is the office chair next to the desk with a monitor?
- Answer: green

https://github.com/ATR-DBI/ScanQA (CVPR 2022)

Tasks & Datasets: Irrelevant QA

RefCOCO dataset

initially used that dataset, which have simple short prompts,
samples from the dataset:



Papers with code. Referring Expression Segmentation on the RefCOCO benchmark .

Tasks & Datasets: Irrelevant QA

RefCOCO dataset

Explored **6 state-of-the-art models** : HIPIE, GLEE-Pro, UniLSeg, PolyFormer, UNINEXT, VLT

Found out that all leading models are not robust to **'negative prompts'**

GLEE [CVPR 2024]

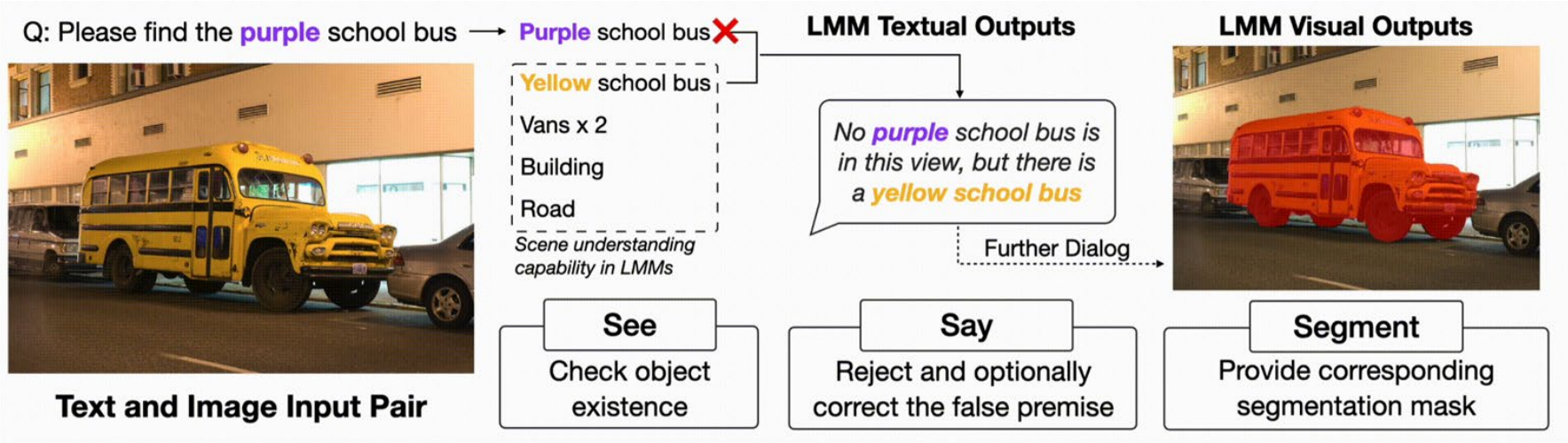
HIPIE [NeurIPS 2023]

UniLSeg [CVPR 2024]

Papers with code. Referring Expression Segmentation on the RefCOCO benchmark .

Tasks & Datasets: Irrelevant QA

Using refCOCO for base images, authors employ an LLM to augment a false-premise referring segmentation dataset with **context -aware false premise queries**



Tsung-Han Wu et al. See, Say, and Segment: Teaching LMMs to Overcome False Premises. CVPR 2024

Tasks & Datasets: SceneVerse (2024)

Scene Caption

Sub-graph Context

{ 'scene_type': 'Bedroom',
'object_count': {'nightstand':2, ...},
'relation': {'nightstand', 'on', 'floor'},
{'backback', 'in front of', bed}, ...}

3D Sub-graph

Summary

Prompt: Provide a summary for a scene from a given scene graph delimited by triple backticks, …
Response: In this bedroom, there are two nightstands, ... The backpack is in front of the nightstand as well. The room appears to be functional, with the nightstands providing storage space and the telephone for communication.

Object Caption

BLIP2 Captions

1. A bed in a hotel room. (0.85)
2. A white comforter on a bed. (0.83)
3. A bed with a striped comforter. (0.83)
…
N. A picture of cat. (0.63)

Multiview Images

Summary

Prompt: Summarize the captions below. The summary should be a description of the {object}. Focus on the {object}'s attributes, like color, shape, material, etc.
Identify and correct the potential errors …
Response: The bed is in a hotel room with a striped comforter. It has a white comforter and a blanket on it. The bed is also in a room with a bedside table.

Object Referral

Relationship Triplets

1. {'table', 'chair', 'left'},
2. {'bed', ('lamp', 'mini fridge'), 'between'}

Template-based Referral

1. The table is to the left of the chair.
2. It's a bed in the middle of a lamp and the mini fridge.

Rephrasing

Prompt: Rewrite the following sentence using one random sentence structure. Focus on the location and relationships about the {target_object}, …
Response:
1. The table is situated to the left of the armchair.
2. The bed occupies the space between the lamp and the mini fridge, creating a cozy atmosphere.

https://scene-verse.github.io/ (ECCV 2024)

Tasks & Datasets: Summary

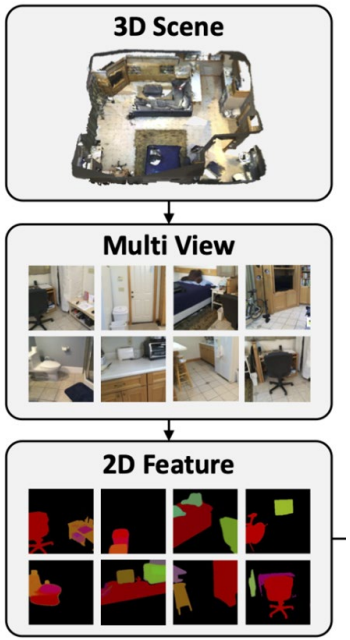
Tasks & Datasets: Summary

- 3D object **captioning** (3D **classification)**
- 3D **detection**
- 3D **semantic** , 3D **instance** , 3D **panoptic** segmentation
- 3D **visual grounding** as a development of the 3D detection task
- 3D **Question Answering**

3

3D-LLM, SpatialRGPT

VLLMs: 3D-LLM (first LLM for 3D)

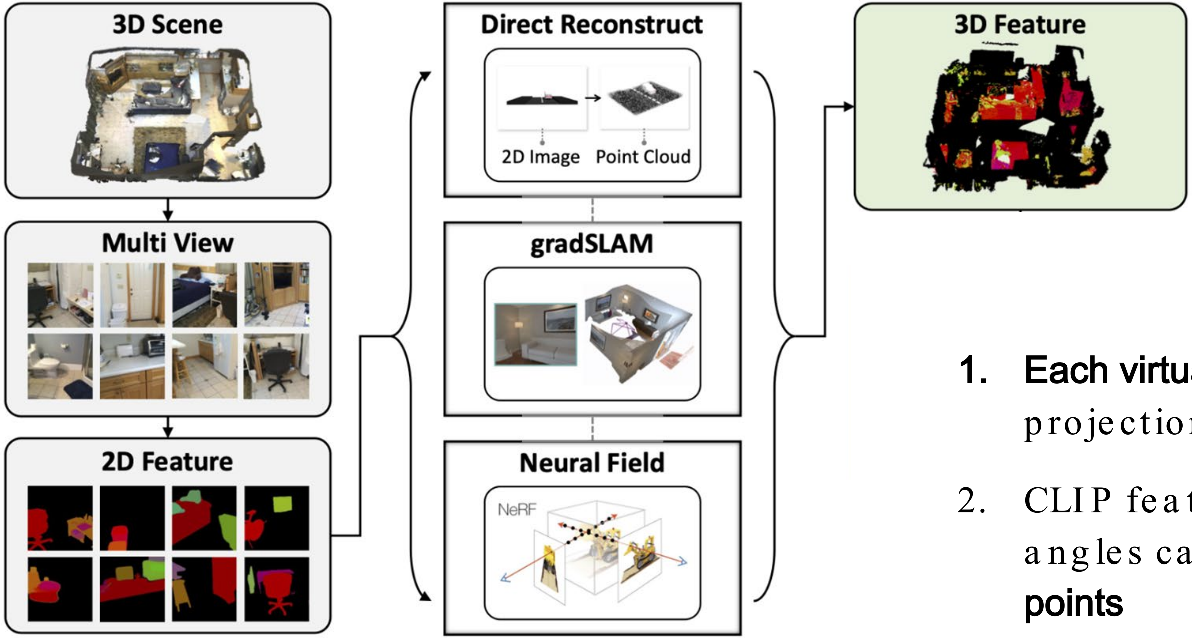


CLIP features are **meaningful for 2D case!**

1. Take our **3D scene** (a cloud of dots or a mesh of a room).
2. Using virtual cameras, we **gather photos of the scene** from different angles
3. For each render, we run CLIP and get a feature vector

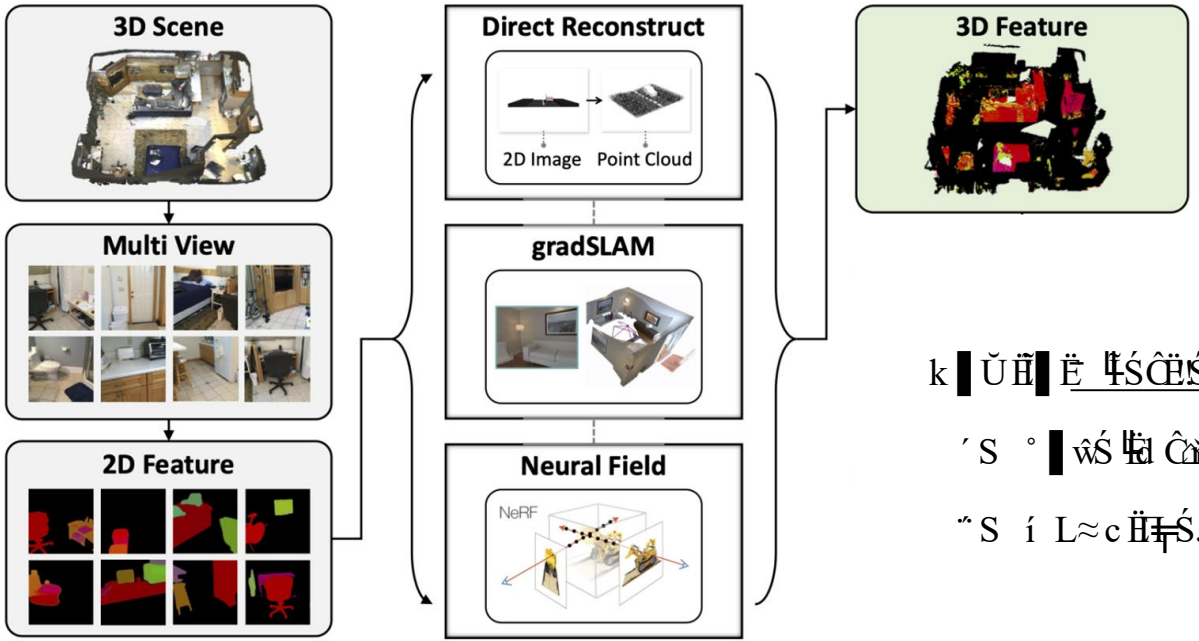
https://github.com/UMass-Embodied-AGI/3D-LLM (NeurIPS 2023)

VLLMs: 3D-LLM (first LLM for 3D)



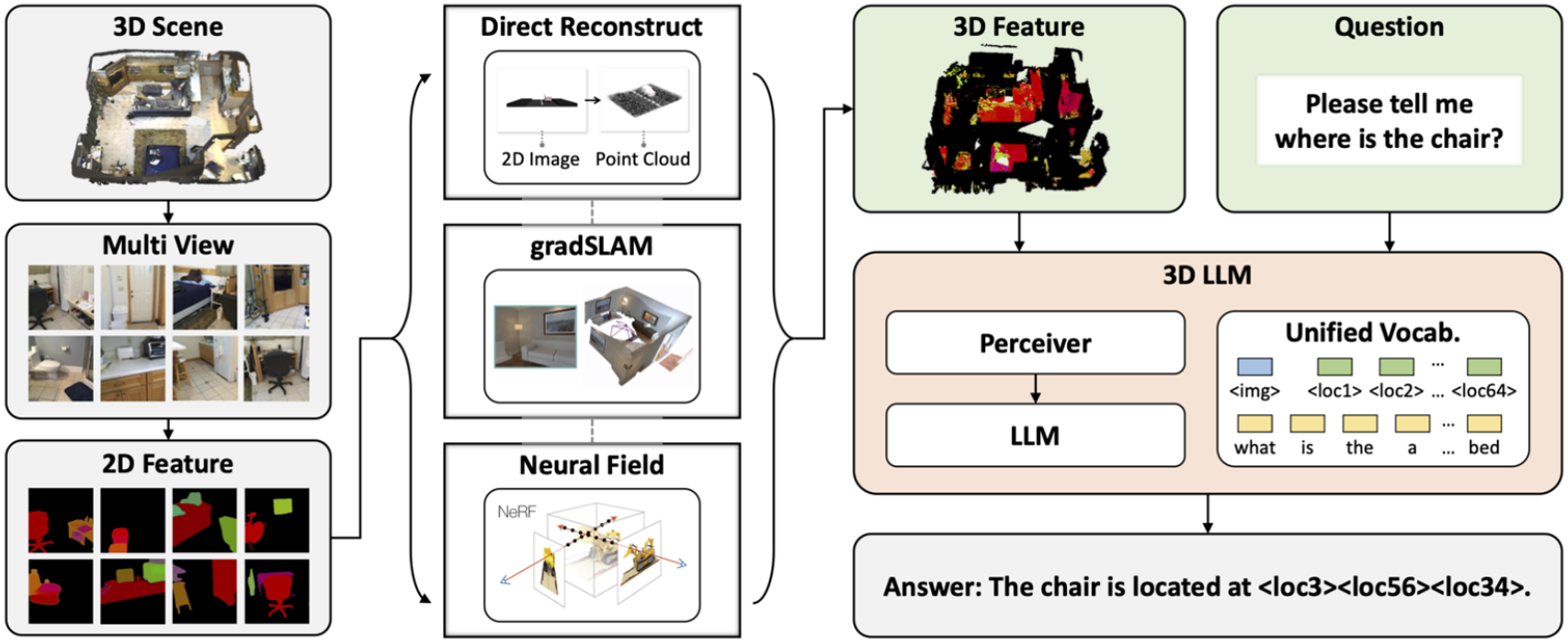
1. **Each virtual camera** has its own projection matrix and depth map
2. CLIP features from different angles can be "sewn" to **3D points**

VLLMs: 3D-LLM (first LLM for 3D)



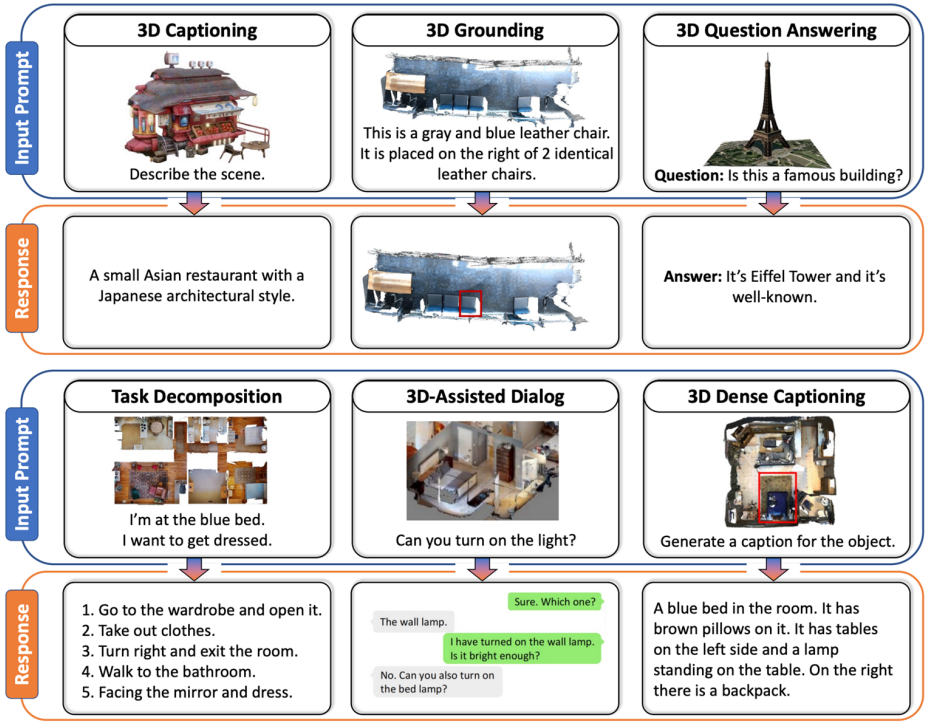
https://github.com/UMass-Embodied-AGI/3D-LLM (NeurIPS 2023)

VLLMs: 3D-LLM (first LLM for 3D)



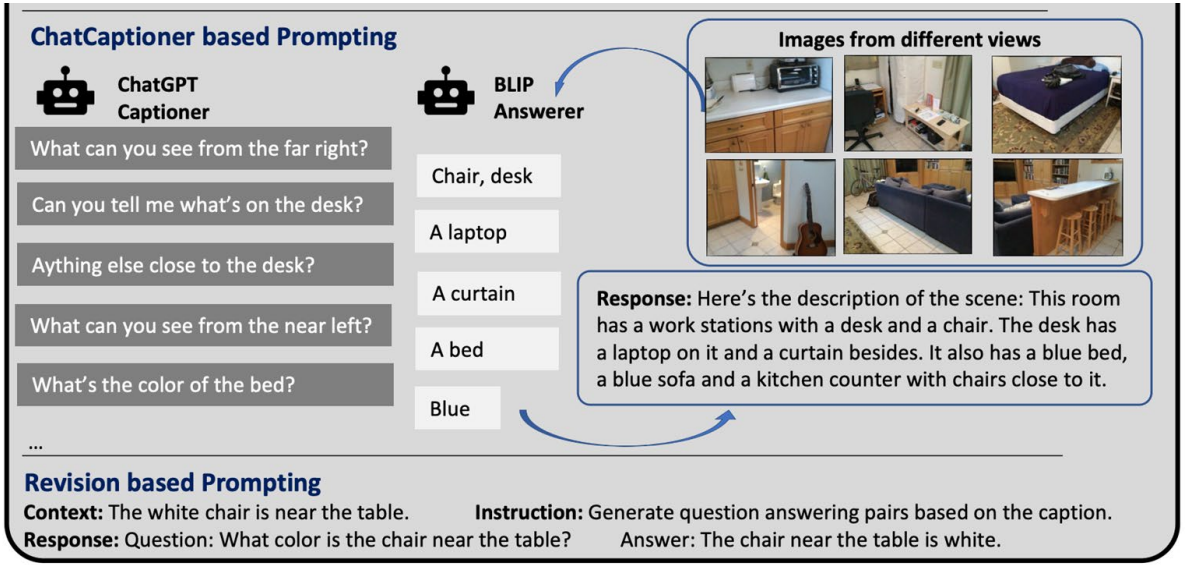
https://github.com/UMass-Embodied-AGI/3D-LLM (NeurIPS 2023)

3D-LLM: Data Generation



- ÎĦ↕½ÎĦ◢▌ŚIË╚▌ ₫ÎŚňËń½Î½ĖŚÎĖË▌ ĈË °ƒ ËĦ╒▌ŬĦŬŚĖ↕½ŭŚĦĈĦĽŚ╤ŚB▌▌ĈËPË
- °₫ĖĴ½Ë▌▌▌ňŚ╚ĖŬ▌ĜŋŚňĦŬ₫ÎĖ╒ńŭŭńĴ½Ë◢▌½ⁿ ŚĖIË╒▌ÎËŭŚĈŶĦŬŚ╙
- d¬€3E° ↕½ĖË╒▌ÎË⁄₀ ◦ Ś½ĈŚňËŶŚÎ

3D-LLM: Data Generation



3D-language data generation pipelines

- Generate questions with ChatGPT, answers with **BLIP -2,** aggregate questions and rewrite them
- Make more diverse dataset from **ScanQA**

https://github.com/UMass-Embodied-AGI/3D-LLM (NeurIPS 2023)

3D-LLM: Results

	B-1	B-2	B-3	B-4	METEOR	ROUHE-L	CIDER	EM
VoteNet+MCAN*	28.0	16.7	10.8	6.2	11.4	29.8	54.7	17.3
ScanRefer+MCAN*	26.9	16.6	11.6	7.9	11.5	30	55.4	18.6
ScanQA*	30.2	20.4	15.1	10.1	13.1	33.3	64.9	**21.0**
LLaVA(zero-shot)	7.1	2.6	0.9	0.3	10.5	12.3	5.7	0.0
flamingo-SingleImage	23.8	14.5	9.2	8.5	10.7	29.6	52	16.9
flamingo-MultiView	25.6	15.2	9.2	8.4	11.3	31.1	55	18.8
BLIP2-flant5-SingleImage	28.6	15.1	9.0	5.1	10.6	25.8	42.6	13.3
BLIP2-flant5-MultiView	29.7	16.2	9.8	5.9	11.3	26.6	45.7	13.6
3D-LLM (flamingo)	30.3	17.8	12.0	7.2	12.2	32.3	59.2	20.4
3D-LLM (BLIP2-opt)	35.9	22.5	16.0	9.4	13.8	34.0	63.8	19.3
3D-LLM (BLIP2-flant5)	**39.3**	**25.2**	**18.4**	**12.0**	**14.5**	**35.7**	**69.4**	20.5

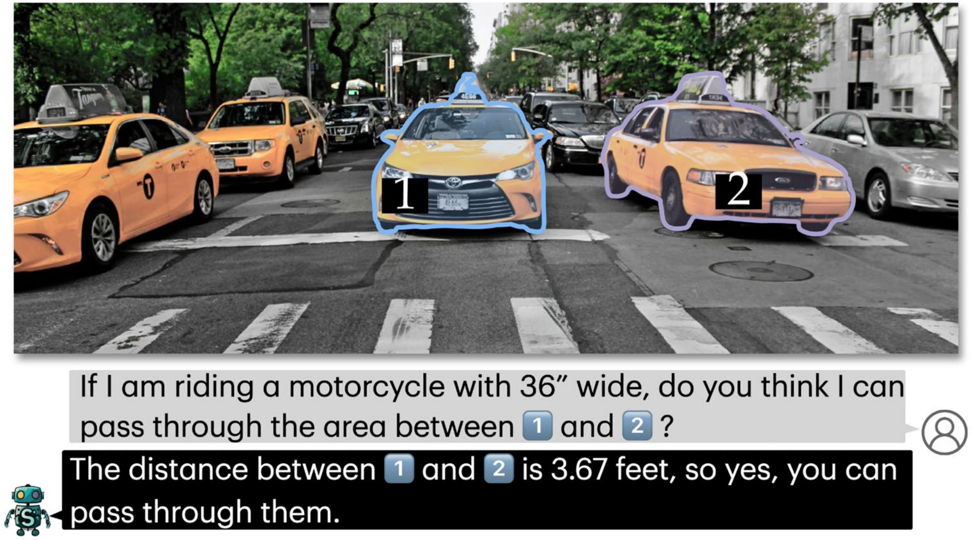
results on **ScanQA** validation set

- Then, **2D models** in zero-shot gave poor results for 3D tasks
- **LLM gives a boost** in the 3D QA task
- 3D features are important

https://github.com/UMass-Embodied-AGI/3D-LLM (NeurIPS 2023)

VLLMs: SpatialRGPT (nvidia)

Go beyond **ScanNet-style datasets** (which are mostly indoor 3D scans)

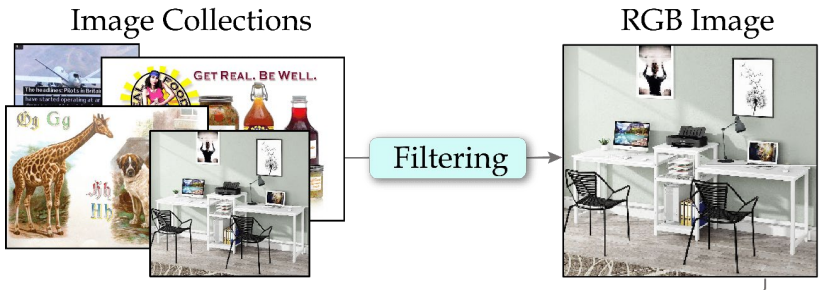


SpatialRGPT is a Vision-Language Model (VLM) focused on **spatial reasoning**

SpatialRGPT works **purely on 2D images** but learns to understand 3D spatial relationships

VLLMs: SpatialRGPT

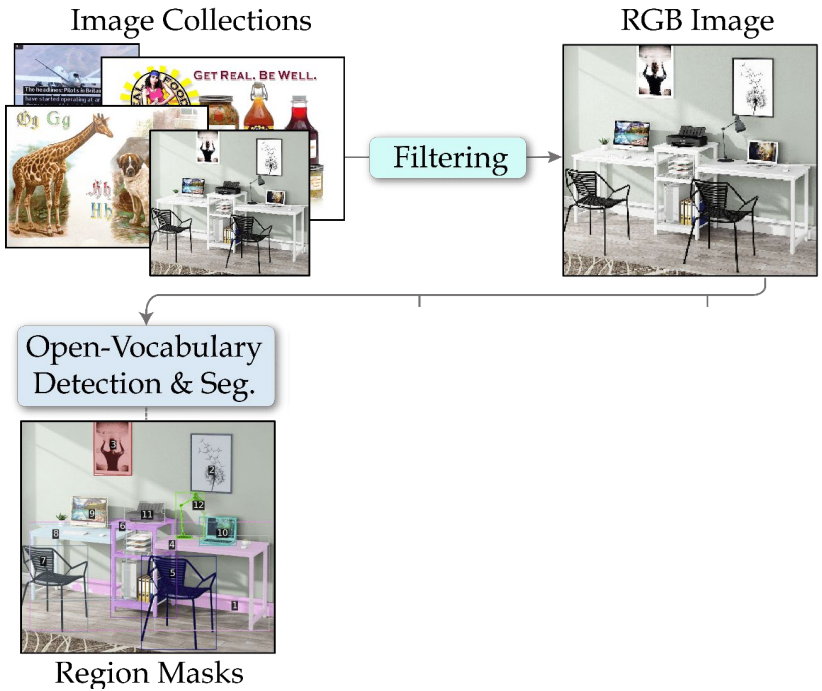
½Ĵ Ĩ▌╢ ½Ĩ ₫'Ëñ ½Ĩ½Ë'Ĵ Ĉ½Ĩ ₫▌╤Ë ⌂ Ś ╚╤Ś



Filtering process to remove any unsuitable images

VLLMs: SpatialRGPT

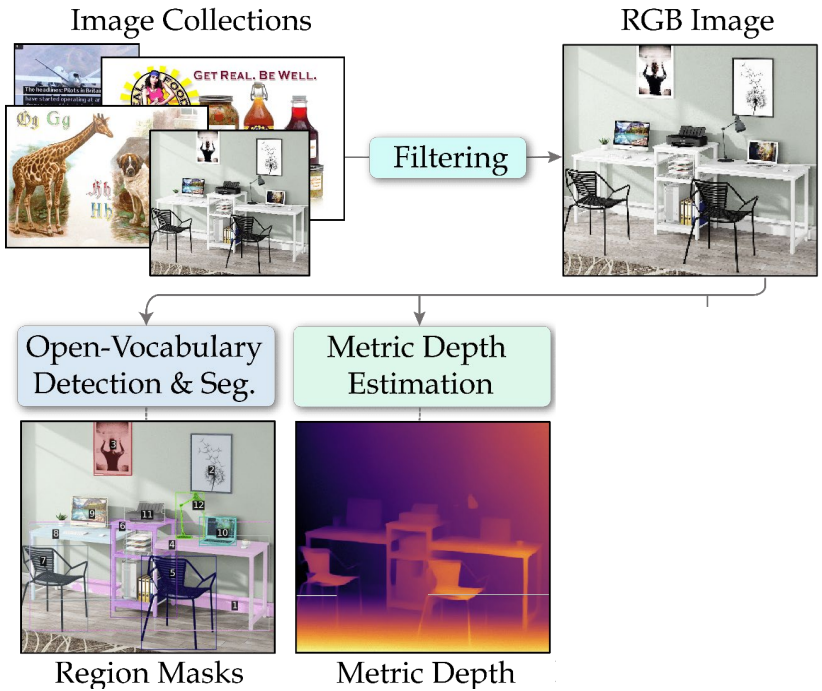
automatic data curation pipeline



- **GroundingDINO** model to detect object bounding boxes

VLLMs: SpatialRGPT

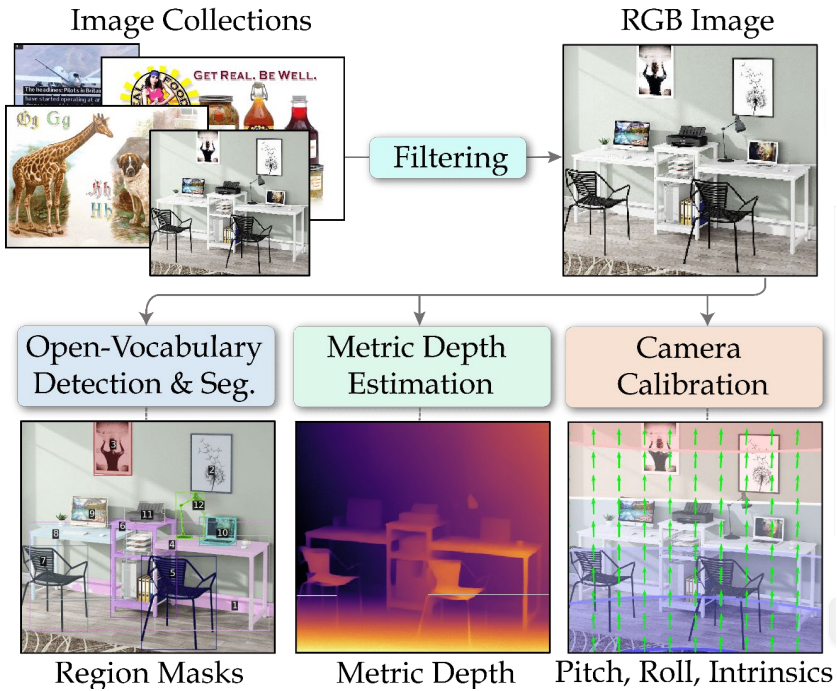
automatic data curation pipeline



- **GroundingDINO** model to detect object bounding boxes
- Estimate metric depth of every pixel using **Metric3Dv2** (focal length)

VLLMs: SpatialRGPT

automatic data curation pipeline

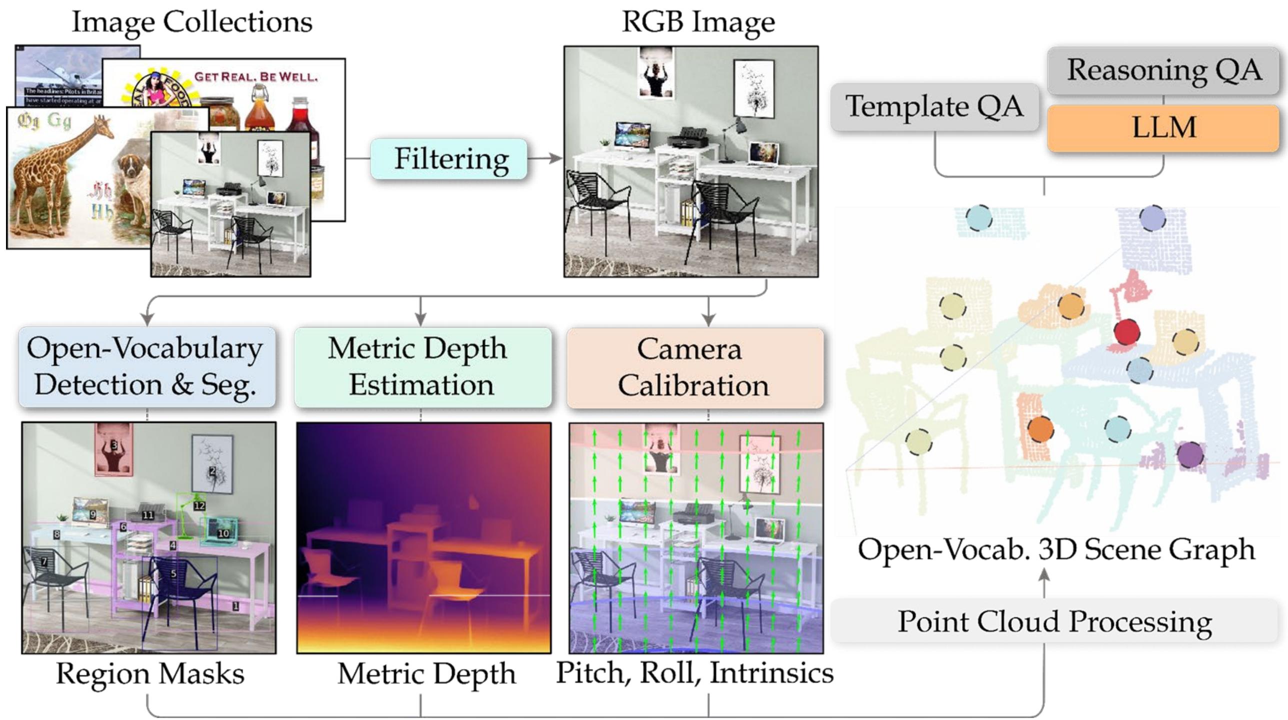


- Calibrate camera intrinsics with **WildCamera** and **PerspectiveFields**

$$K = \begin{pmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix}$$

Œ ½Ĩ⌂½╘d ¬€ĬĒû ∘ Ś╤ĬŒ ½Ĩ⌂½╬B ½Ĩ½ĖŚĨ

automatic data curation pipeline



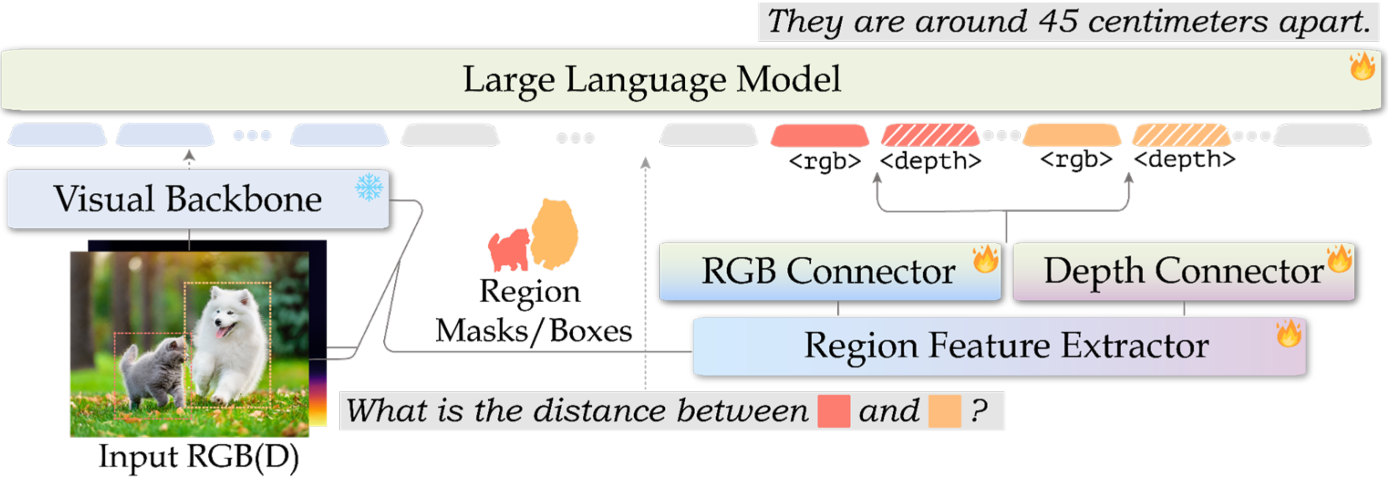
Project each object mask into **3D space** using depth + camera info

Derive a **scene graph**

Make **rich QA pairs** , ~8.7M QA concepts

VLLMs: SpatialRGPT

RoI-Align for generated masks, ViT processes the raw **RGB image** into a grid of feature tokens, **depth map** passes through a small CNN



VLLMs: SpatialRGPT

How far is 0 from 1 horizontally?

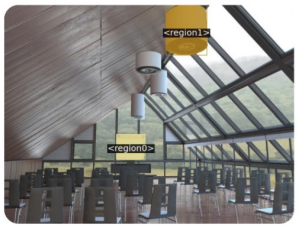
0 and 1 are 31.21 feet apart horizontally.

Could you tell me the vertical size of 0?

0 is 3.67 feet tall.

Does 0 have a larger size compared to 1?

Correct, 0 is larger in size than 1.



Does 0 have a lesser width compared to 1?

No, 0 is not thinner than 1.

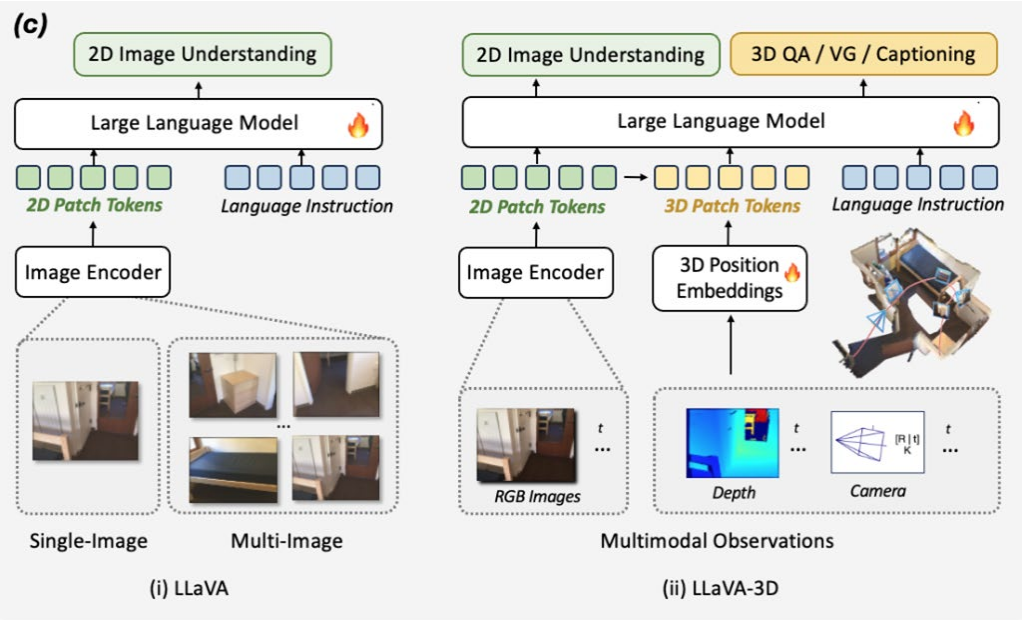
First spatial dataset, for testing authors used part of their benchmark

Method	Below/ Above	Left/ Right	Big/ Small	Tall/ Short	Wide/ Thin	Behind/ Front	Avg.
GPT-4	64.1	42.8	42.8	61.6	61.6	49.0	57.8
GPT-4V	63.3	46.6	64.1	60.7	68.2	45.4	58.1
LLaVA-v1.6-34B	44.1	45.7	36.7	53.5	37.5	45.4	43.9
GPT-4V+SoM	75.0	55.2	42.4	54.4	49.0	47.2	54.3
LLaVA-v1.6-34B+SoM	44.1	40.0	33.9	47.3	41.3	46.3	42.3
Kosmos-2	28.3	15.2	4.71	26.7	12.5	12.7	17.0
RegionVILA	30.8	47.6	35.8	44.6	35.5	49.0	40.4
SpatialRGPT	**99.1**	**99.0**	79.2	89.2	83.6	87.2	89.8
SpatialRGPT-Depth	**99.1**	**99.0**	**80.1**	**91.9**	**87.5**	**91.8**	**91.7**

SpatialRGPT-Bench Qualitative Results: Numbers represent success rates in percentage (↑).

https://www.anjiecheng.me/SpatialRGPT

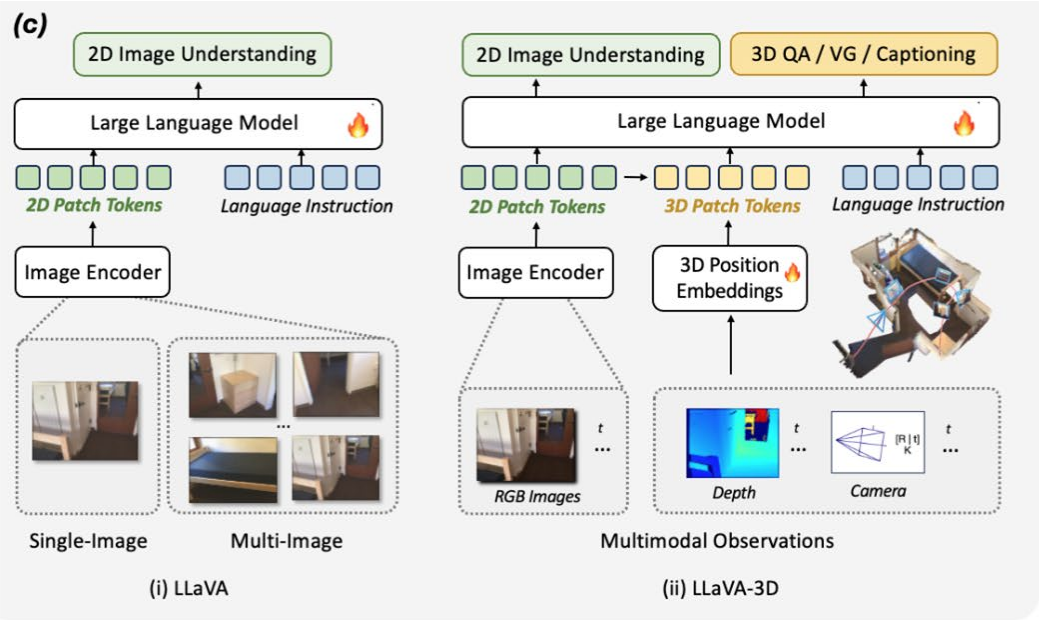
VLLMs: LLaVA-3D

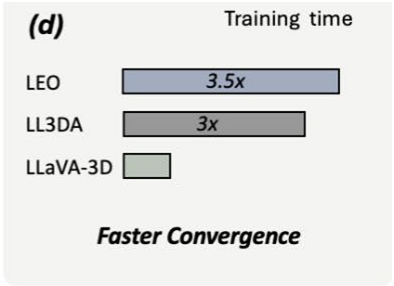


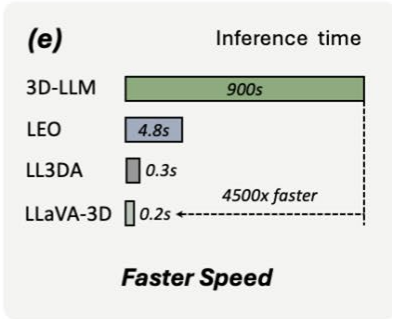
- Preserves the ability to understand **2D images**
- Expand the model with the part connected with the **3D understanding**
- Give the images from **different angles**

https://zcmax.github.io/projects/LLaVA-3D/ (2024-09)

VLLMs: LLaVA-3D







ê Ĵ L’↕ Ë!½ËĨŚĈË
Ĩ↕½╤ËAB3ÜÜê ĨË
ŁŚËËËŚ½ĨĴ ĈŚËIË
Śŵ° Ś╤Ė▫ŚËAB Ë
ĈŚŁ▌╤ËĨĈĴ LĨ┛
╤

VLLMs: LLaVA-3D

Generated their **own dataset** (combination of 2D and 3D) – 1M instructions, **~sota**

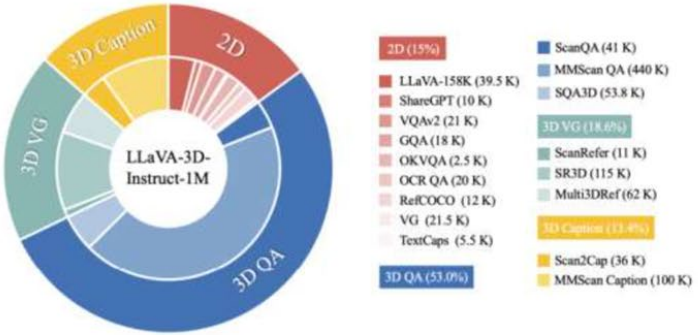


Figure 3. **LLaVA-3D-Instruct-1M.** The hybrid 2D and 3D Dataset Collection. Left: Distribution of data across categories, with the outer circle representing all categories and the inner circle illustrating data subset distribution. Right: Detailed dataset quantities.

	ScanQA (val)					SQA3D (test)
	C	B-4	M	R	EM@1	EM@1
Task-specific models						
Scan2Cap [11]	-	-	-	-	-	41.0†
ScanRefer+MCAN [49]	55.4	7.9	11.5	30.0	18.6	-
ClipBERT [25]	-	-	-	-	-	43.3
ScanQA [3]	64.9	10.1	13.1	33.3	21.1	47.2
3D-VisTA [53]	69.6	10.4	13.9	35.7	22.4	48.5
Task-specific fine-tuned 3D LMMs						
3D-LLM (FlanT5) [16]	69.4	12.0	14.5	35.7	20.5	
LL3DA [35]	76.8	13.5	15.9	37.3	-	
Chat-3D v2 [17]	87.6	14.0	-	-	-	54.7
LEO [18]	**101.4**	13.2	20.0	49.2	24.5 (47.6)	50.0 (52.4)
Scene-LLM [15]	80	12.0	16.6	40.0	**27.2**	54.2
Zero-shot 2D LMMs						
VideoChat2 [30]	49.2	9.6	9.5	28.2	19.2	37.3
LLaVA-NeXT-Video [26]	46.2	9.8	9.1	27.8	18.7	34.2
GPT-4V	59.6	-	13.5	33.4	-	-
Gemini	68.3	-	11.3	35.4	-	-
Claude	57.7	-	10.0	29.3	-	-
LLaVA-3D	91.7	**14.5**	**20.7**	**50.1**	27.0 (45.0)	**55.6** (57.6)

https://zcmax.github.io/projects/LLaVA-3D/ (2024-09)

4

; B3 ň Ś

3D-LLM, SpatialRGPT

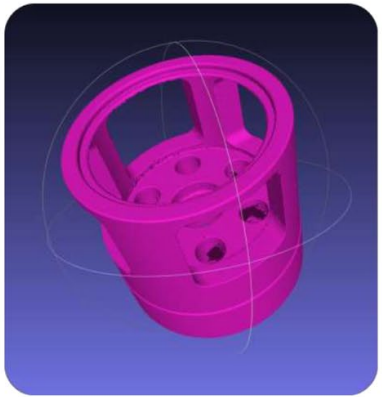
CAD-models: Reverse Engineering

Build a full-fledged **constructive model (CAD)** of the selected part in the engineering software, in order to produce an exact copy of the part

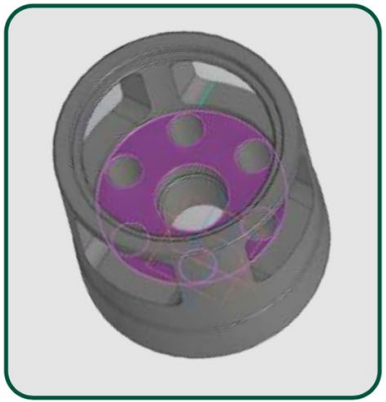
CAD-models: Reverse Engineering

Mesh approximation (piecewise linear approximation of a cylinder) ⟶ It can be done using **splines** — boundary-representation (exact mathematical model)

photo



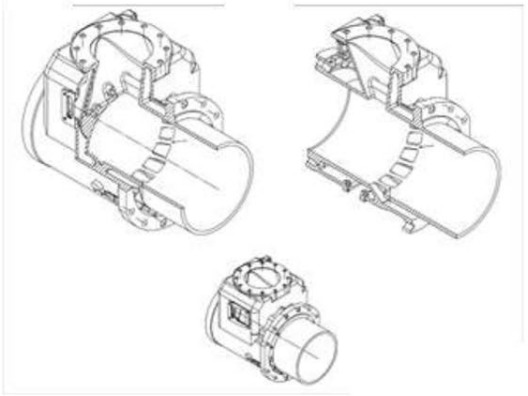
ŚĖ↕Ë



B-rep

CAD-models: Reverse Engineering

Mesh approximation (piecewise linear approximation of a cylinder) ⟶ It can be done using **splines** — boundary-representation (exact mathematical model)



real device with multiple details

ÁBË▌ňŚ

; B3╣ ▌ňŚ╚ËÏË≈ŚūŚĈËŚËĽ╤ⁿ ◁╤ŚŚĈ◁╤ⁿ Ë

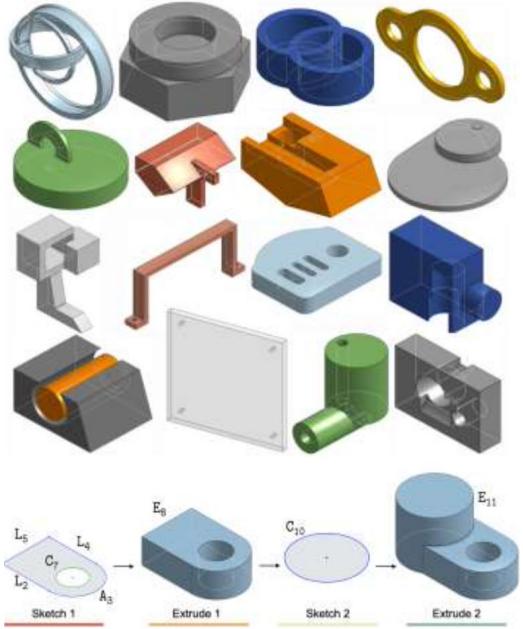
real device with multiple details

1. device is sequentially **disassembled** into parts, **record** the process
2. **calibration** marks + **3D** scanner
3. **manual measurements** of individual elements
4. scan (**mesh**) of the model is loaded into the software
5. **engineer manually designs** the part so that the result matches the scan
6. **manager validates** the result

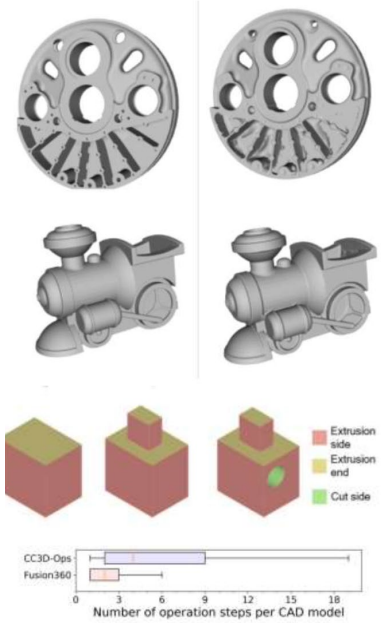
CAD-models: Reverse Engineering

The part can be significantly **damaged**

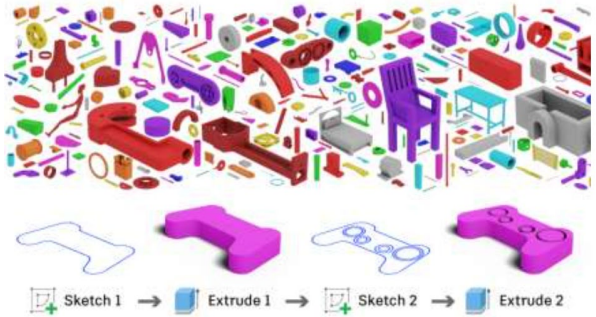
CAD-models: Benchmarks



BŚŚ◦ ; B



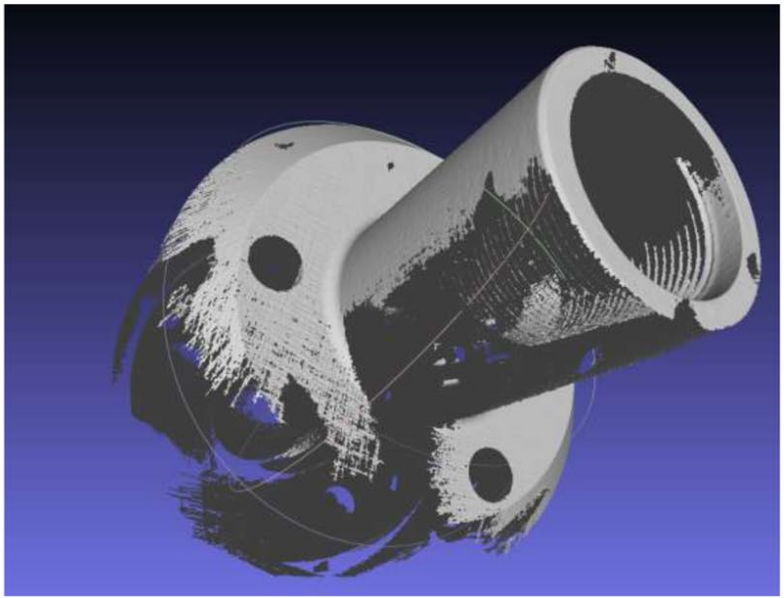
cvi2

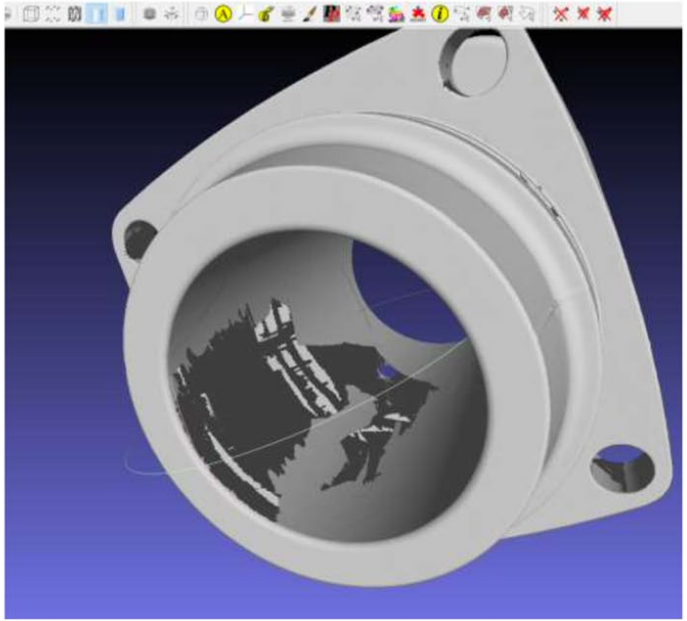


AutodeskAILab

CAD-models: Benchmarks

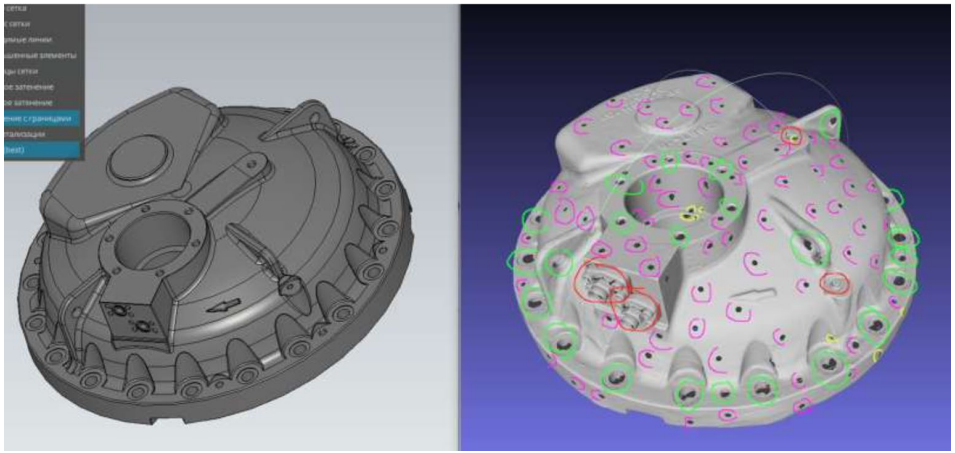
There are **no surface pieces in the scan** due to the limited scanning angles in practice





CAD-models: Benchmarks

1. **random holes** in the mesh (need to be repaired)
2. round holes in the mesh due to **marks** (need to be repaired)
3. **round holes in the part** itself that look almost like round holes in the mesh (can not be repaired)
4. rart of the mesh belongs to **another part**



Many more artifacts

CAD-models: Benchmarks

Этап 1: Формообразующие операции

1. **Выдавливание (Extrude)**
2. **Вращение (Revolve)**
3. **Элемент по траектории (Sweep)**
4. **Элемент по сечениям (Loft)**
5. **Булевы операции (Boolean)**
6. **Обечайка (Shell)**
7. **“Операция без истории”**
8. **Отдельные детали по ГОСТу**

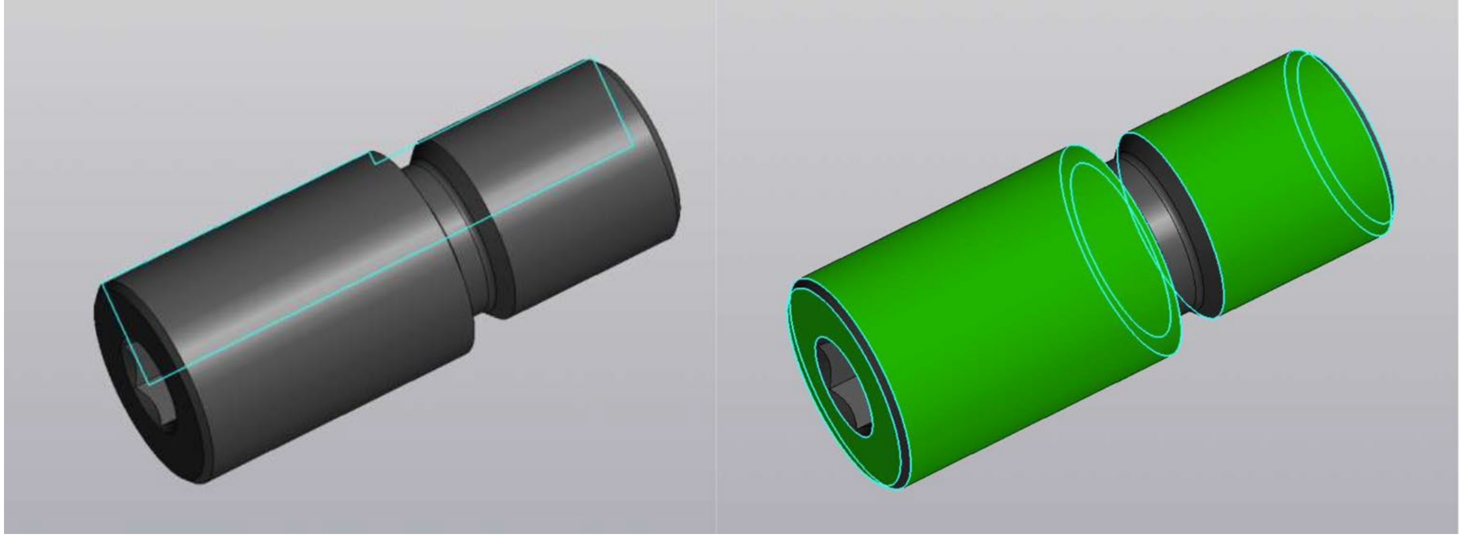
- **очень часто**
- **довольно часто**
- **редко**
- **один раз**

Этап 2: Дорабатывающие операции

1. **Отверстие (Hole)**
2. **Фаски (Chamfer)**
3. **Скругления (Fillet)**
4. **Массивы (Array, Pattern)**
5. **Надписи / гравировки (Text)**
6. **Резьба**
7. **Прочие детали оформления: “отверстие ГОСТ 14034-74” (Model 30.7), условные обозначения и т. п.**

CAD-models: Benchmarks

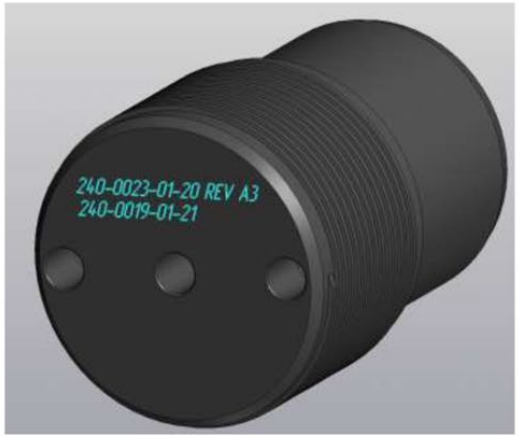
Вращение как формообразующая операция



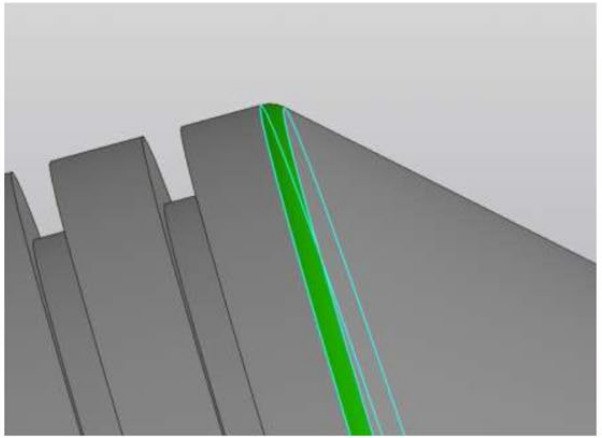
Винт сердечника

CAD-models: Benchmarks

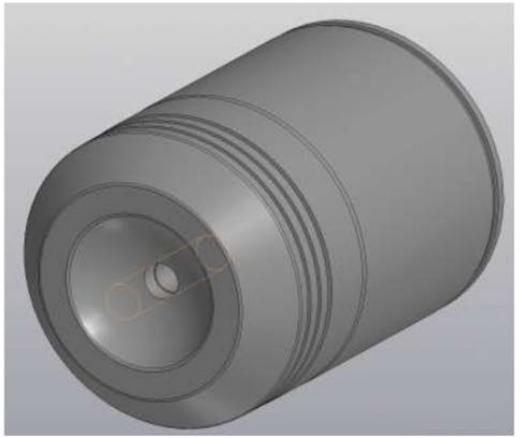
Примеры дорабатывающих операций



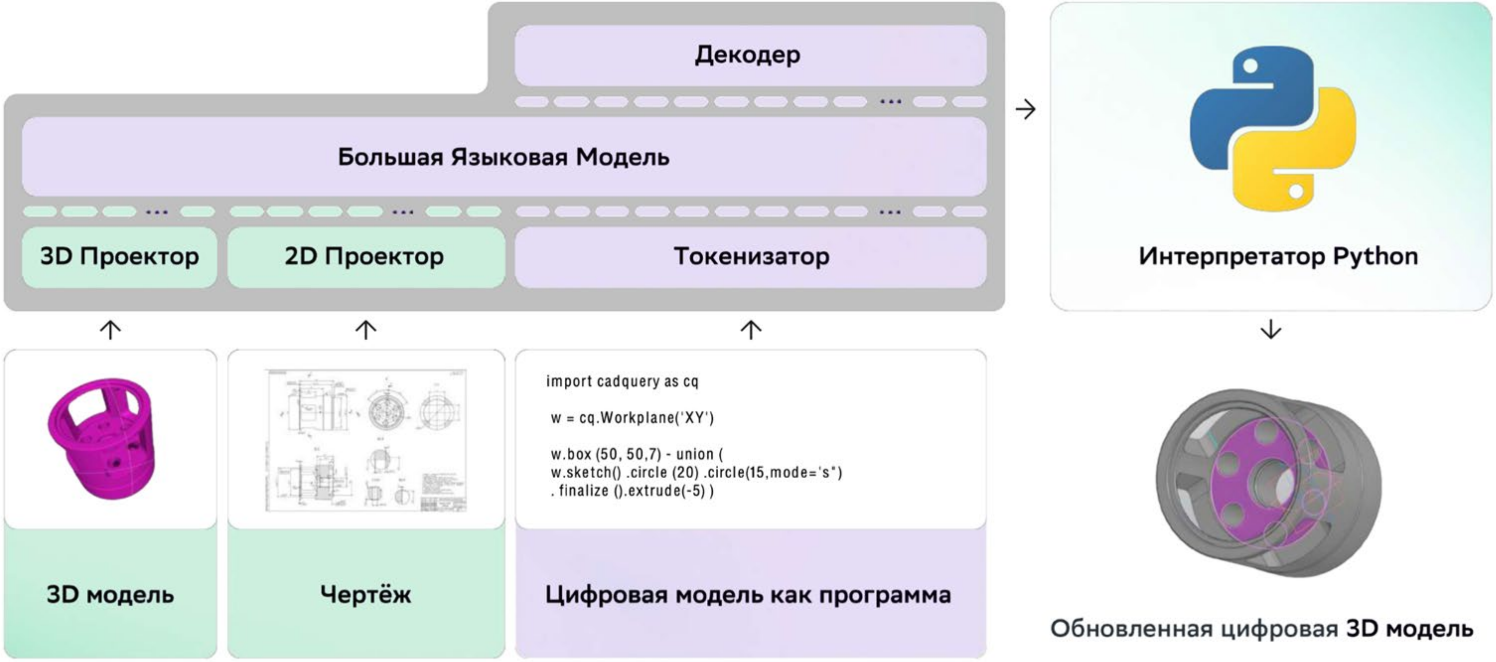
Надпись



Скругление

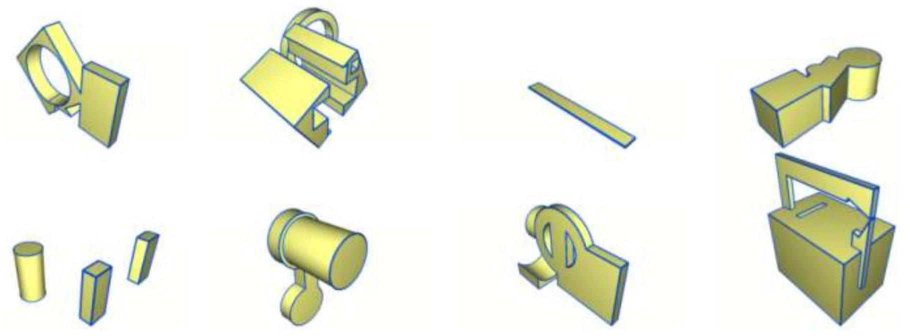


Отверстие



CAD-Recode: Reverse Engineering CAD Code from Point Clouds

CAD-models: Datasets

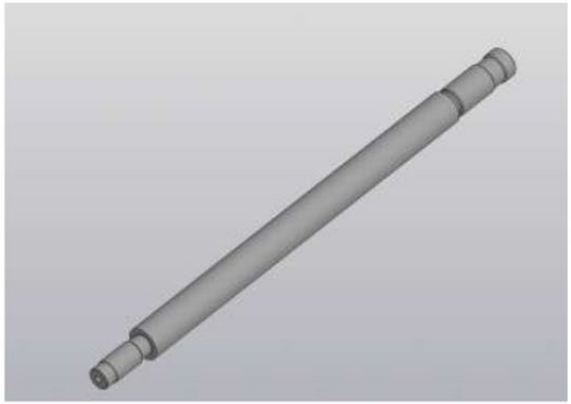
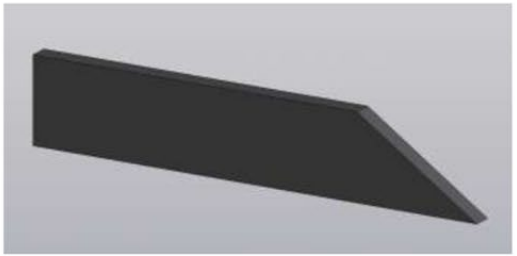
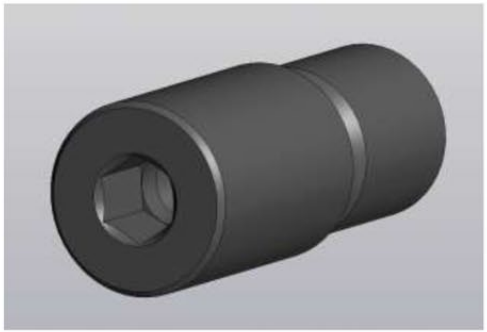


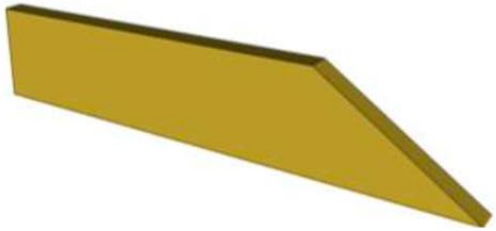
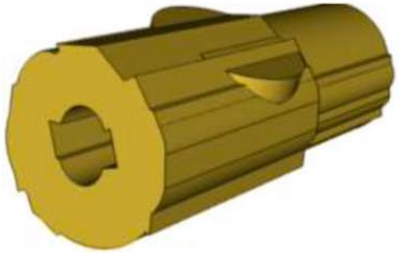
LLM-based

- Randomized generation of examples based on operations (profile, extrude)
- The examples are not very meaningful

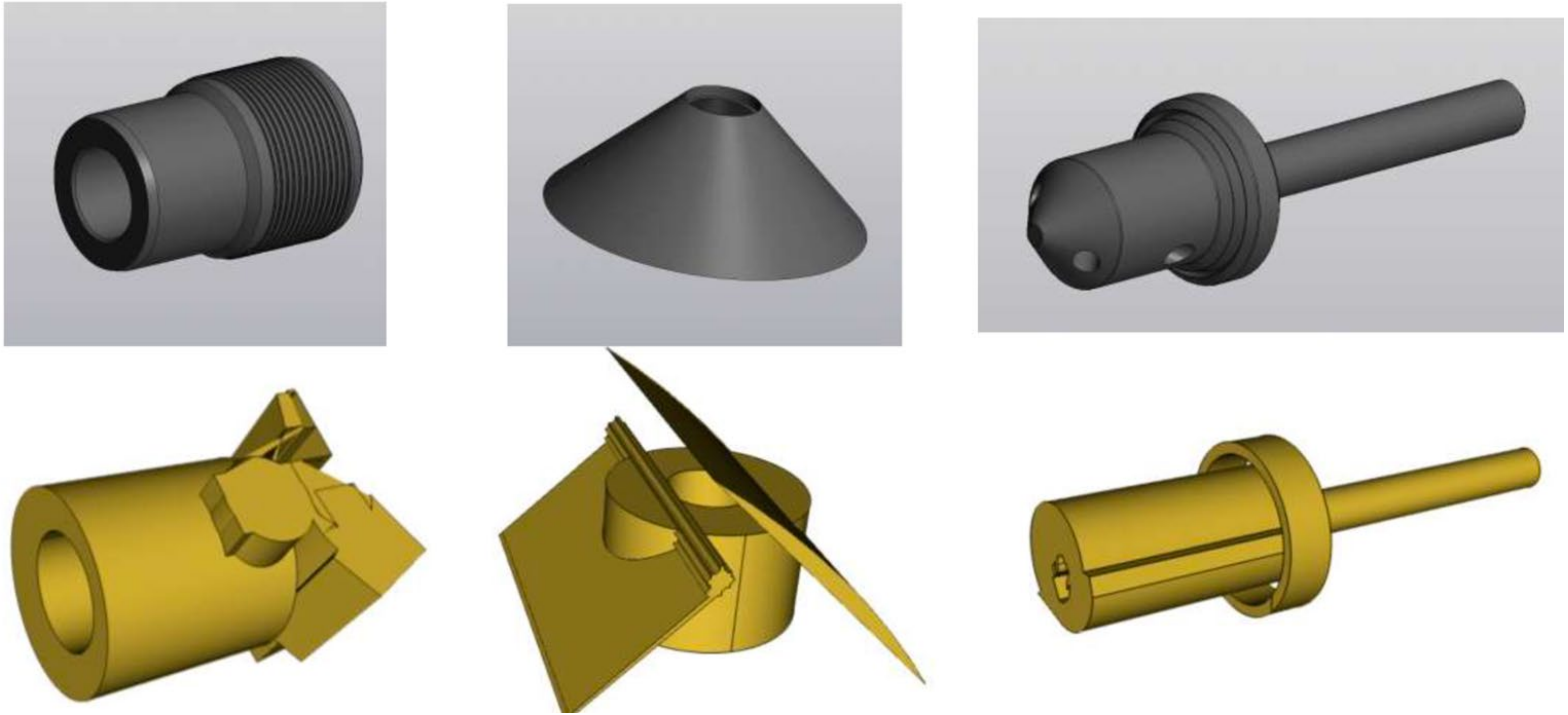
- Identify 10-20 common types of parts
- Use LLM to get the parameterization of the parts based on the description of the part class

CAD-models: MLLMs

; B3╢ ▌ňŚ ╚ËïË ÜÜê ĖË



Conclusions

1. **3D vision** = semantic + metric visions

2. c▌ĈË ½Ĩ↕ŚĈ╕ⁿ ËABËñ½Ĩ½ ŬŚËĽ½╤ĤĴ ĖŚË
ňŚ◦ Ĩ↕ËĽ½╢ ŚĈ½ĖIËÜ◿B ≈ĖIËę▌◹╤ĨËĽ┤ĴňĖ

3. ≈ŚĽŚ╤ĨË½ňūŁ½╤ĽŚ╢ Ś╤Ĩ▌!ËABËê ÜÜê Ï
ňŁ½Ĩ½ĖŚĨËę ĈŚ◦ ½Ĉ½ĨĴ▌╤ËьËABËŚ½ĨĴ ĈŚË
Ś╤Ľ▌ňŁ╕ ËĽ▌ ËĽ▌╖Ś╤ĖË

Когда прочитал весь материал на курсе, который смог, и студенты пошли сдавать бизнес-проекты