Fast SceneScript: Fast and Accurate Language-based 3D Scene Understanding via Multi-Token Prediction

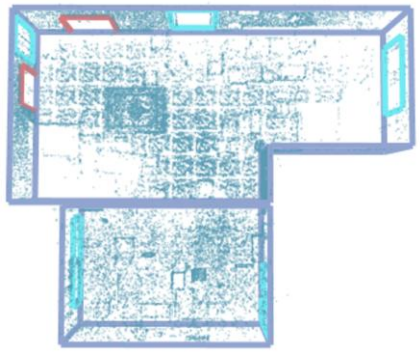
Ruihong Yin[1, 2] Xuepeng Shi[1] Oleksandr Bailo[1] Marco Manfredi[1] Theo Gevers[2]

[1]Qualcomm XR Labs

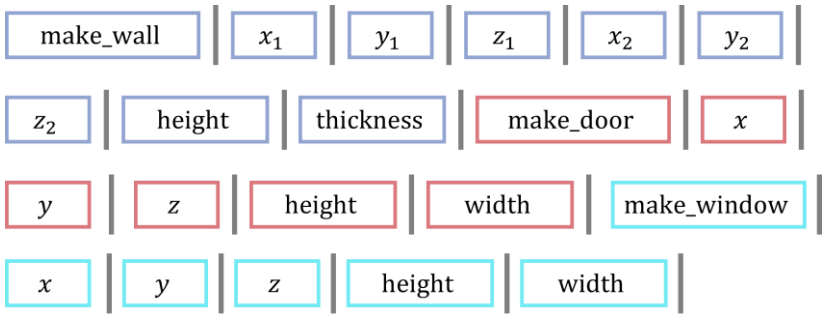
[2]University of Amsterdam

Introduction

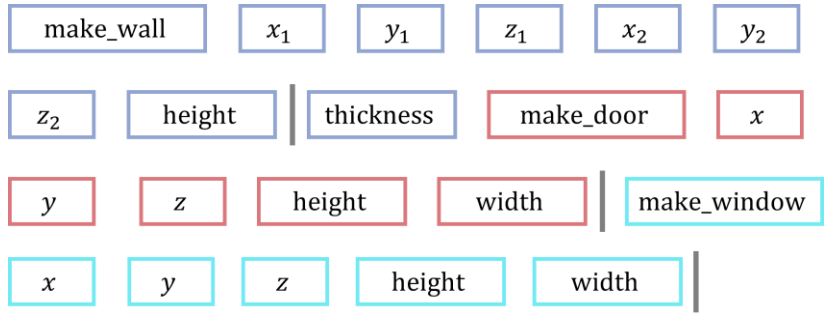
- In the unified language-based perception approaches, next-token prediction (NTP) causes **high latency** particularly when the sequence length increases.
- Unlike natural language, the structured language designed for 3D perception is **more deterministic and weakly coupled**.



(a) 3D scene layout



(b) Inference pricess by SceneScript (**21** iterations)



(c) Inference pricess by our Fast SceneScript (**3** iterations)

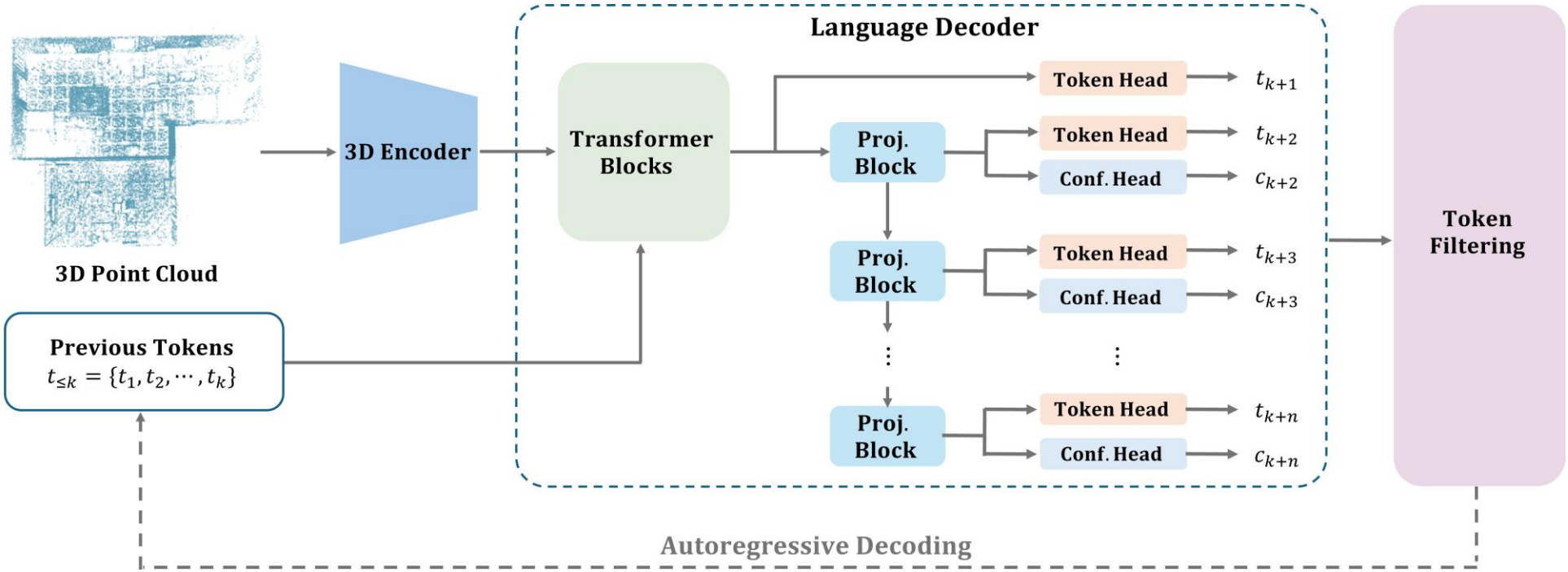
Method - make the unified model **fast** and **accurate**

- Employ **multi-token prediction** (MTP) to accelerate the inference
  - 5x speedup compared to SceneScript

- Investigate the **decoding strategies with token filtering mechanisms** to achieve accurate and reliable inference
  - 12% improvement in F1-Score compared to vanilla MTP

- Design a **parameter-efficient mechanism** to reduce the number of parameters
  - 43% few parameters compared to vanilla MTP

Multi-token prediction (MTP)



Framework of Fast SceneScript

Token Filtering Strategies

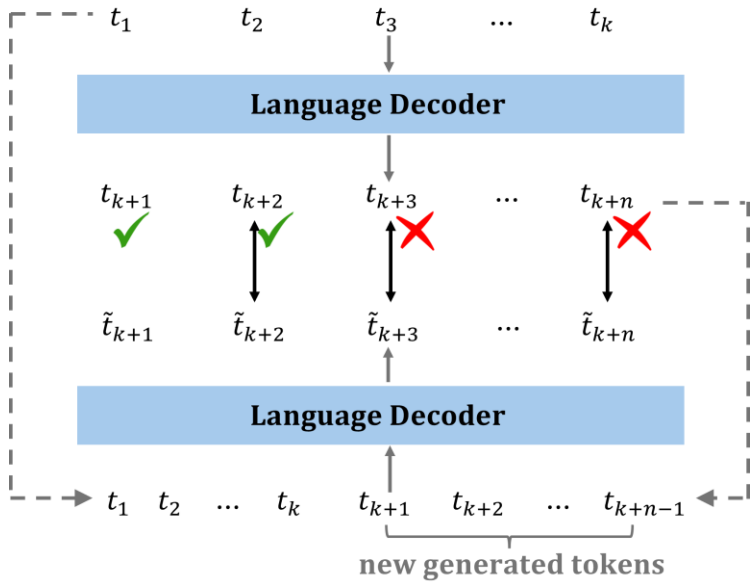
- **Self-Speculative Decoding (SSD):**

Step 1: Draft n future tokens $\{t_{k+1}, t_{k+2}, \cdots, t_{k+n}\}$ by **the *n* token heads**.

Step 2: Generate $\{\tilde{t}_{k+1}, \tilde{t}_{k+2}, \cdots, \tilde{t}_{k+n}\}$ with the **first token head**.

Step3: Assess the alignment between $\{t_{k+2}, \cdots, t_{k+n}\}$ and $\{\tilde{t}_{k+2}, \cdots, \tilde{t}_{k+n}\}$.

Step 4: Only accept tokens before the first unreliable token.



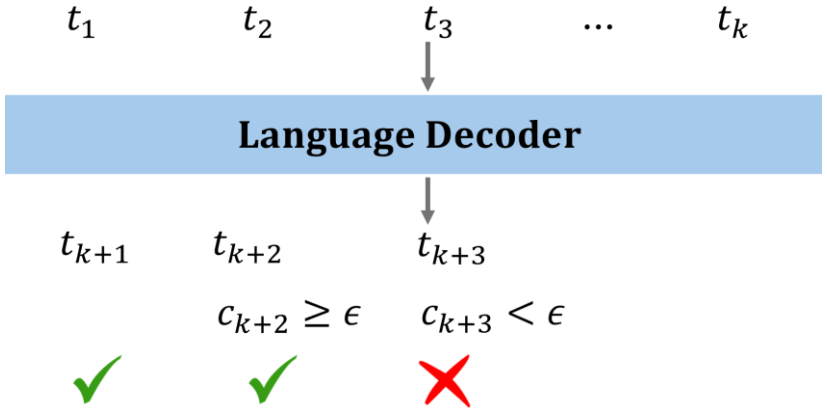
Self-Speculative Decoding (SSD)

- **Confidence-Guided Decoding (CGD):**

Step 1: Draft n future tokens $\{t_{k+1}, t_{k+2}, \cdots, t_{k+n}\}$ and predict their confidences by **the *n* token heads**.

Step 2: Assess the reliability by checking their confidences.

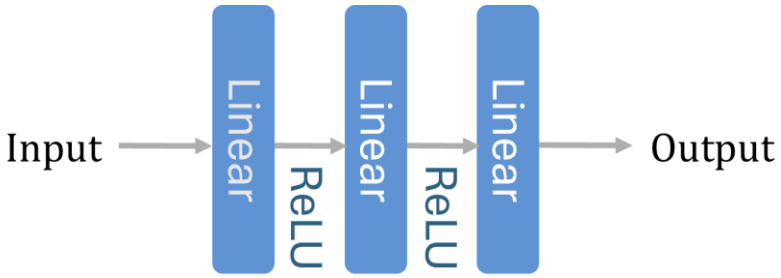
Step 3: Only accept tokens before the first unreliable token.



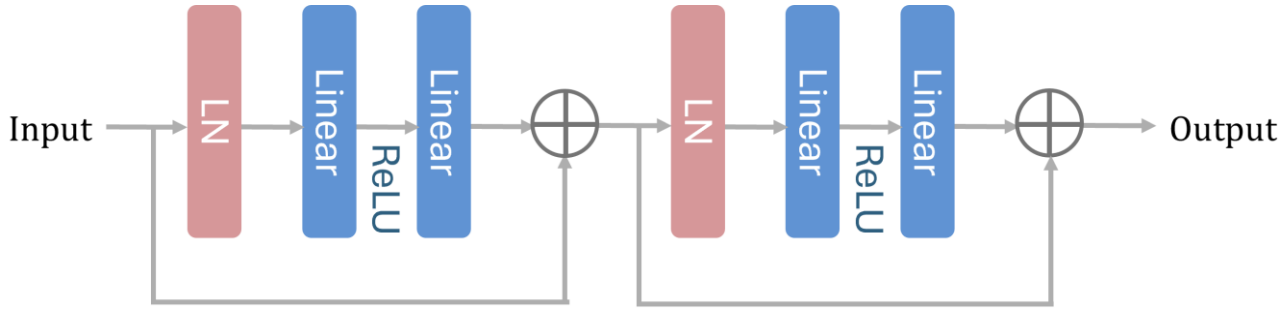
Confidence-Guided Decoding (CGD)

Parameter-Efficient Mechanism

- **Shared token head and confidence head** to reduce the number of parameters.
- **Shared projection block** to generate distinct hidden states for each token.



Token / Confidence Head



Projection Block

Experiments

- **Datasets**:

  Synthetic: ASE, Structured3D

  Real-world: SceneCAD

- **Metrics**:

  Accuracy:

  F1-Score: computed at every 5cm interval from 0m to 1m.

  Efficiency:

  Param: the number of learnable parameters in the language decoder.

  Latency: the average decoding time for predicting layout for one scene.

  α: the average number of accepted tokens per decoder inference.

Experiments

- **Comparisons for layout estimation on ASE dataset**

	Method	n	Param ↓	Latency ↓	α_{val} ↑	F1-Score of *val* set ↑ wall	window	door	mean	α_{test} ↑	F1-Score of *test* set ↑ wall	window	door	mean
a	SceneScript * [2]	1	14.00 M	387 ms	1	0.890	0.865	0.918	0.891	1	0.896	0.865	0.928	0.896
b	SceneScript [2]	1	14.00 M	382 ms	1	0.918	0.880	0.940	0.913	1	0.921	0.881	0.942	0.915
c	SceneScript [2] + MTP [12]	4	18.14 M	109 ms	4	0.891	0.852	0.913	0.885	4	0.898	0.855	0.913	0.889
d	SceneScript [2] + MTP [12]	6	20.91 M	76 ms	6	0.835	0.813	0.880	0.843	6	0.838	0.813	0.889	0.847
e	SceneScript [2] + MTP [12]	8	23.67 M	62 ms	8	0.831	0.804	0.885	0.840	8	0.836	0.804	0.886	0.842
f	Fast SceneScript (SSD)	8	15.05 M	81 ms	7.46	0.914	0.882	0.939	0.912	7.45	0.919	0.882	0.939	0.913
g	Fast SceneScript (CGD)	8	16.10 M	92 ms	6.29	0.912	0.883	0.938	0.911	6.30	0.918	0.883	0.938	0.913
h	SceneScript [2] + MTP [12]	10	26.43 M	54 ms	10	0.805	0.776	0.863	0.815	10	0.808	0.774	0.861	0.814
i	Fast SceneScript (SSD)	10	15.05 M	75 ms	8.97	0.910	0.879	0.937	0.909	8.99	0.915	0.880	0.940	0.912
j	Fast SceneScript (CGD)	10	16.10 M	89 ms	7.27	0.909	0.880	0.936	0.908	7.27	0.912	0.879	0.938	0.910

Experiments

- **Comparisons for object detection on ASE dataset**

	Method	Latency ↓	α ↑	F1 @.25 ↑	F1 @.50 ↑
a	SceneScript [2]	535 ms	1	0.851	0.823
b	SceneScript [2] + MTP [12]	83 ms	8	0.815	0.772
c	Fast SceneScript (SSD)	104 ms	7.16	0.858	0.829
d	Fast SceneScript (CGD)	108 ms	6.53	0.859	0.832

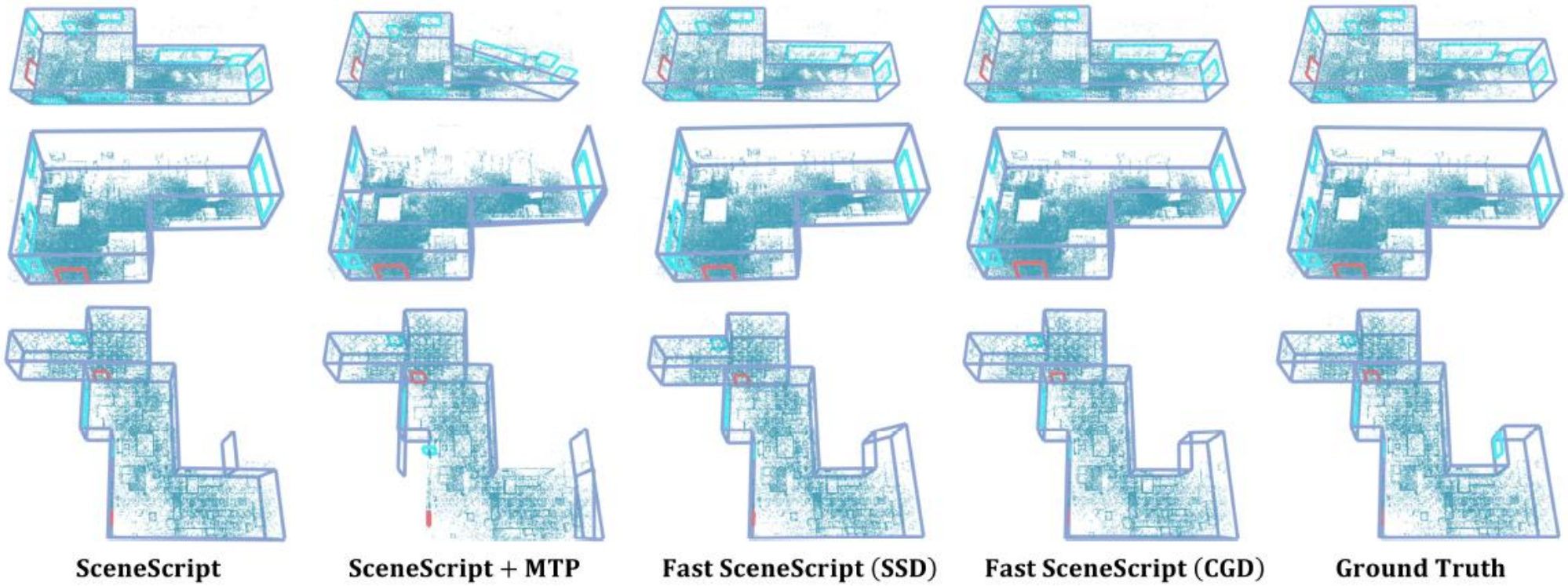
Experiments

- **Comparisons for layout estimation for SceneCAD**

	Method	n	Layout Estimation			Object Detection			
			Latency ↓	α ↑	F1-Score ↑	Latency ↓	α ↑	F1 @.25 ↑	F1 @.50 ↑
a	SceneScript [2]	1	108 ms	1	0.556	104 ms	1	0.670	0.461
b	SceneScript [2] + MTP [14]	8	23 ms	8	0.460	20 ms	8	0.688	0.508
c	Fast SceneScript (SSD)	8	42 ms	5.07	0.556	33 ms	5.72	0.703	0.544
d	Fast SceneScript (CGD)	8	46 ms	4.87	0.552	31 ms	5.68	0.710	0.529

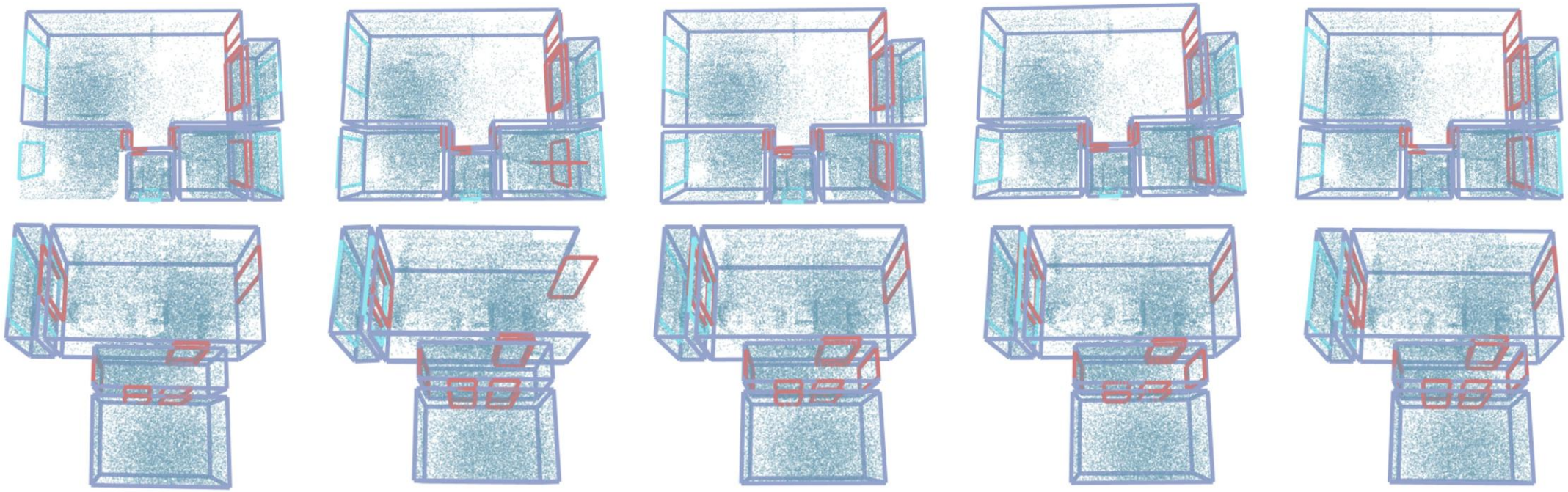
Experiments

- **Visualizations for layout estimation on ASE dataset**



Experiments

- **Visualizations for layout estimation on Structured3D dataset**



Qualcomm UvA

Thanks for watching