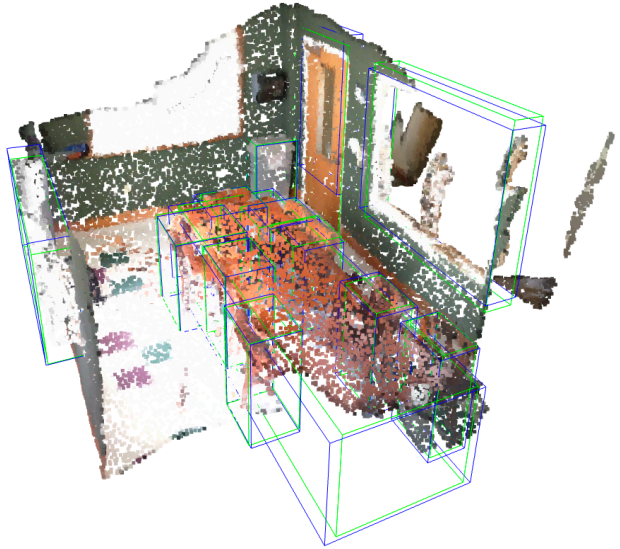
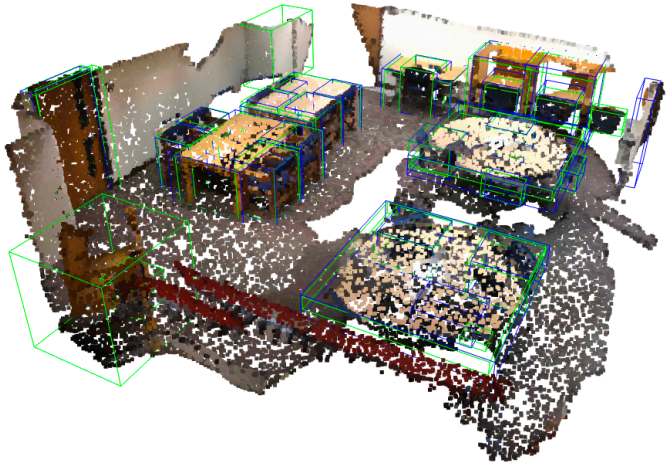
Hierarchical Point Attention for Indoor 3D Object Detection

Manli Shu[1,2] Le Xue[2] Ning Yu[2] Roberto Martín-Martín[2,3]

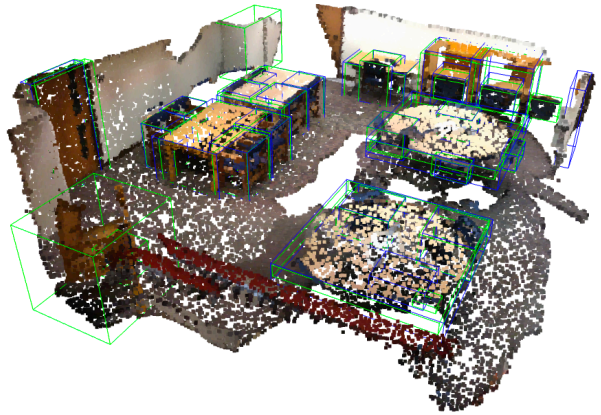
Caiming Xiong[2] Tom Goldstein[1] Juan Carlos Niebles[2,4] and Ran Xu[2]



[1] University of Maryland, College Park, MD, U.S.A.
[2] Salesforce Research, Palo Alto, CA, U.S.A.
[3] University of Texas, Austin, TX, U.S.A.
[4] Stanford University, CA, U.S.A.



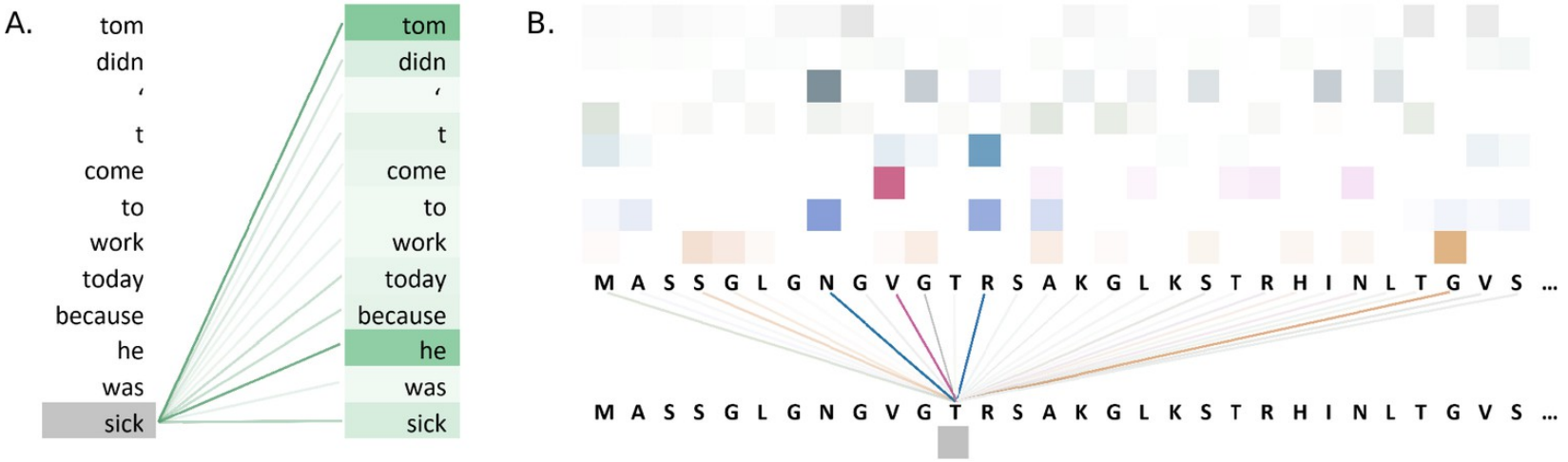
Indoor 3D Object Detection



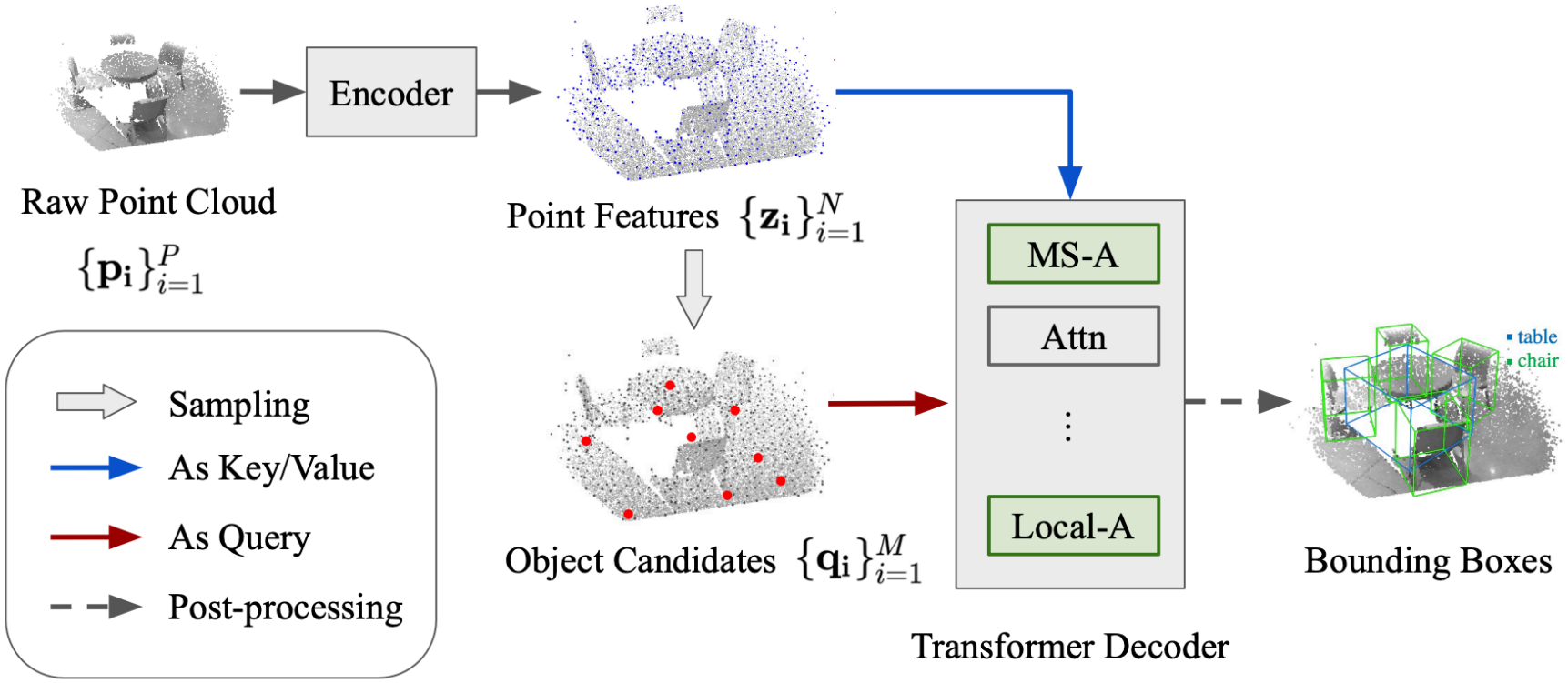
- Applications
  - Augmented Reality
  - 3D room layout planner
  - Home robots
  - …
- Challenges
  - Dense scene
  - Cluttered objects
  - Varying object sizes and shapes

Transformers as 3D Detectors

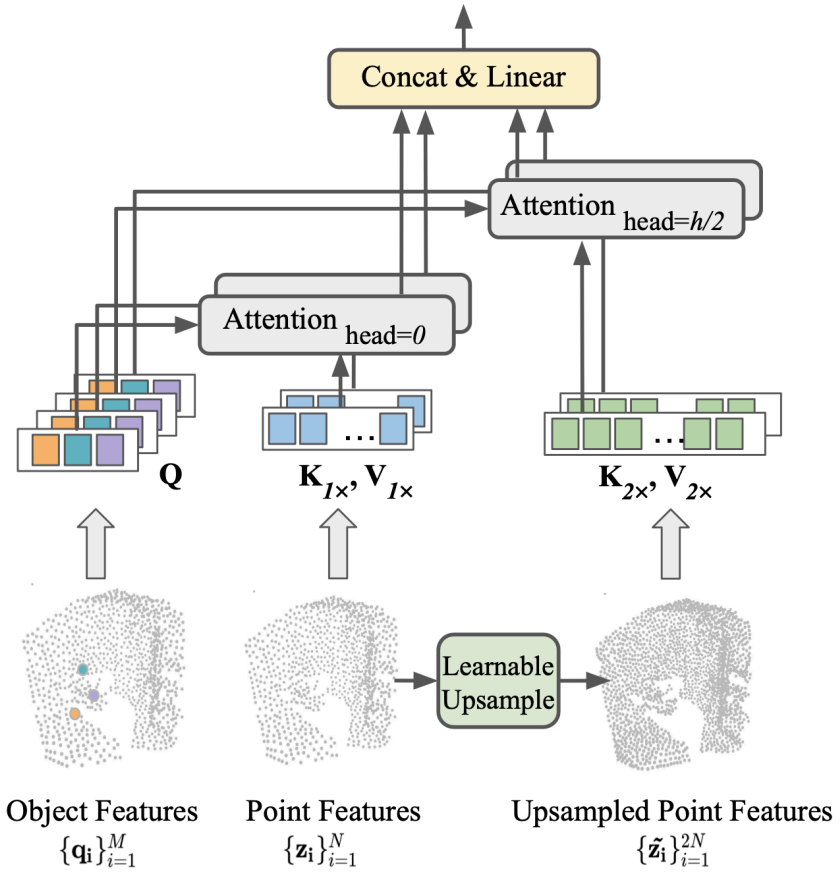
- Transformers do not require input data to have 2d/3d structures
  - Suitable for modeling point cloud data as **a sequence of points**.
- Transformers are good at handling long sequences
  - Modeling **long-range relationships** among all points within a scene
  - Extract rich **global context** information



Point-based 3D transformer detector

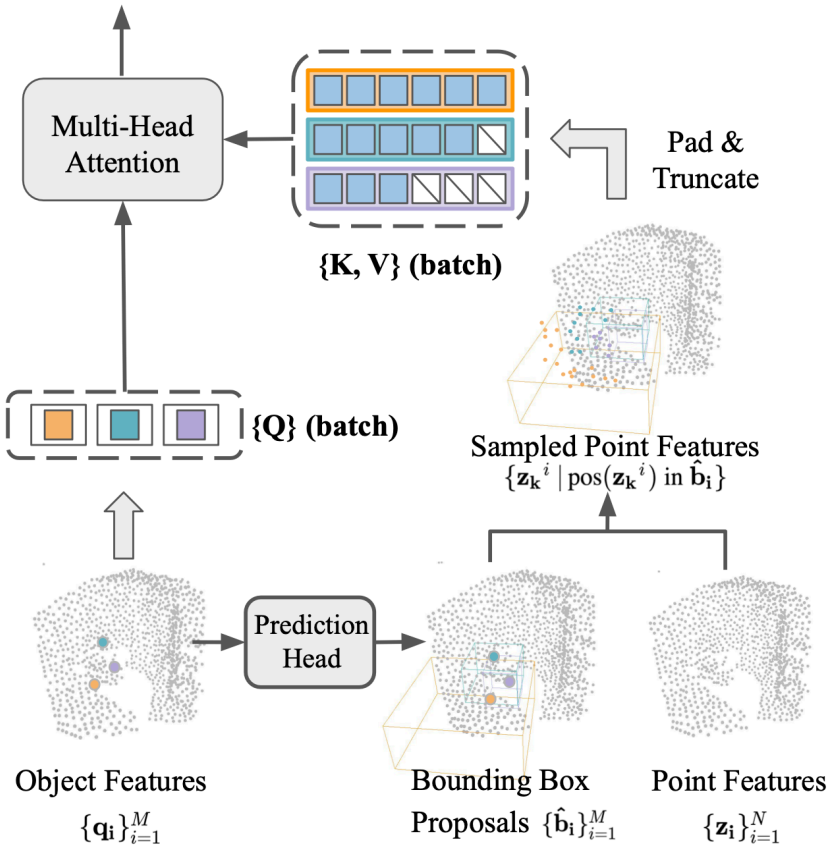


Aggregated Multi-Scale Attention (MS-A)



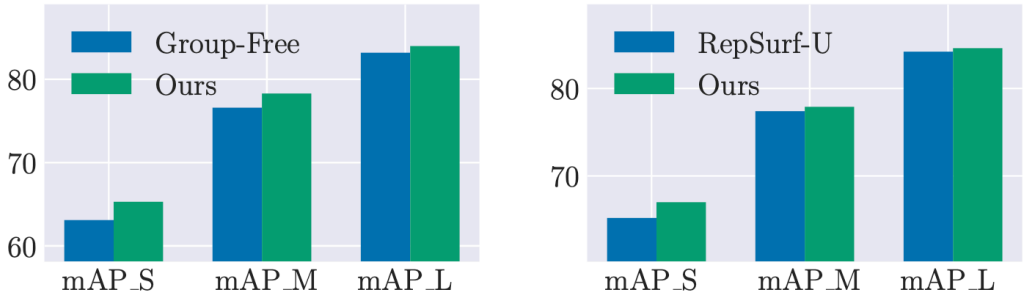
1. Learnable Up-sample
    1. Sample arbitrary points from the raw point clouds
    2. Interpolating the input features to get the initial up-sampled features.
    3. Learnable network layer that projects the interpolated features.
2. Multi-scale feature aggregation
    1. Different subsets of the attention head use features of different resolution.
    2. The forward computation doesn't increase.

Size-Adaptive Local Attention (Local-A)

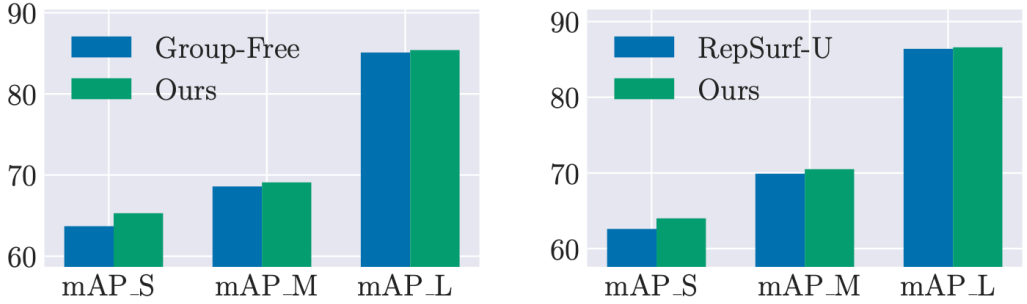


1. Attention region definition
    1. Obtain intermediate bounding box proposals from the current object candidate feature.
    2. For each object (Q), only use point features within its bounding box proposal as the K and V for attention.
2. Batch processing
    1. A batch of object candidates (Q) have different sets of K and V (of different lengths).
    2. We do padding/truncation for batch processing.

Improvements on small objects



(a) Per size-category (S/M/L) mAPs on ScanNetV2.



(b) Per size-category (S/M/L) mAPs on SUN-RGBD.

- We measure the mean average precision (mAP) within different size categories (small / medium / large).
- The proposed hierarchical point attention bring most significant performance gain in small objects.
- Our attention modules can be plugged into any point-based transformer detector.
- We are able to further improve the SOTA model.

Ablation Study

TABLE III

THE EFFECT OF N_{local} IN LOCAL-A.

When there are enough points, a larger N_local means the points are sampled more densely within each bounding box proposal.

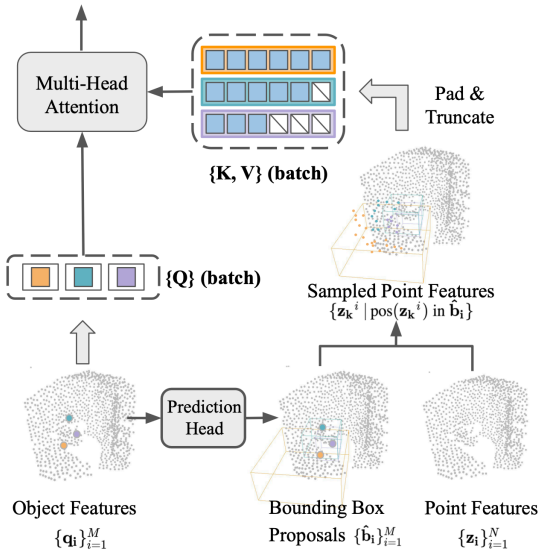
N_{local}	mAP@0.25	mAP@0.50	mAP_S	mAP_M	mAP_L
8	67.8	51.1	64.3	77.2	82.8
16	**68.8**	**52.3**	65.1	**77.9**	83.4
24	68.7	**52.3**	**65.2**	77.7	83.5
32	68.3	52.1	64.7	77.3	**83.8**

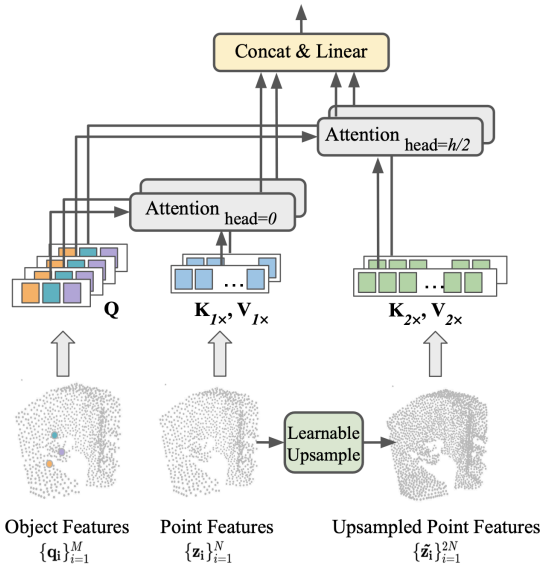
TABLE IV

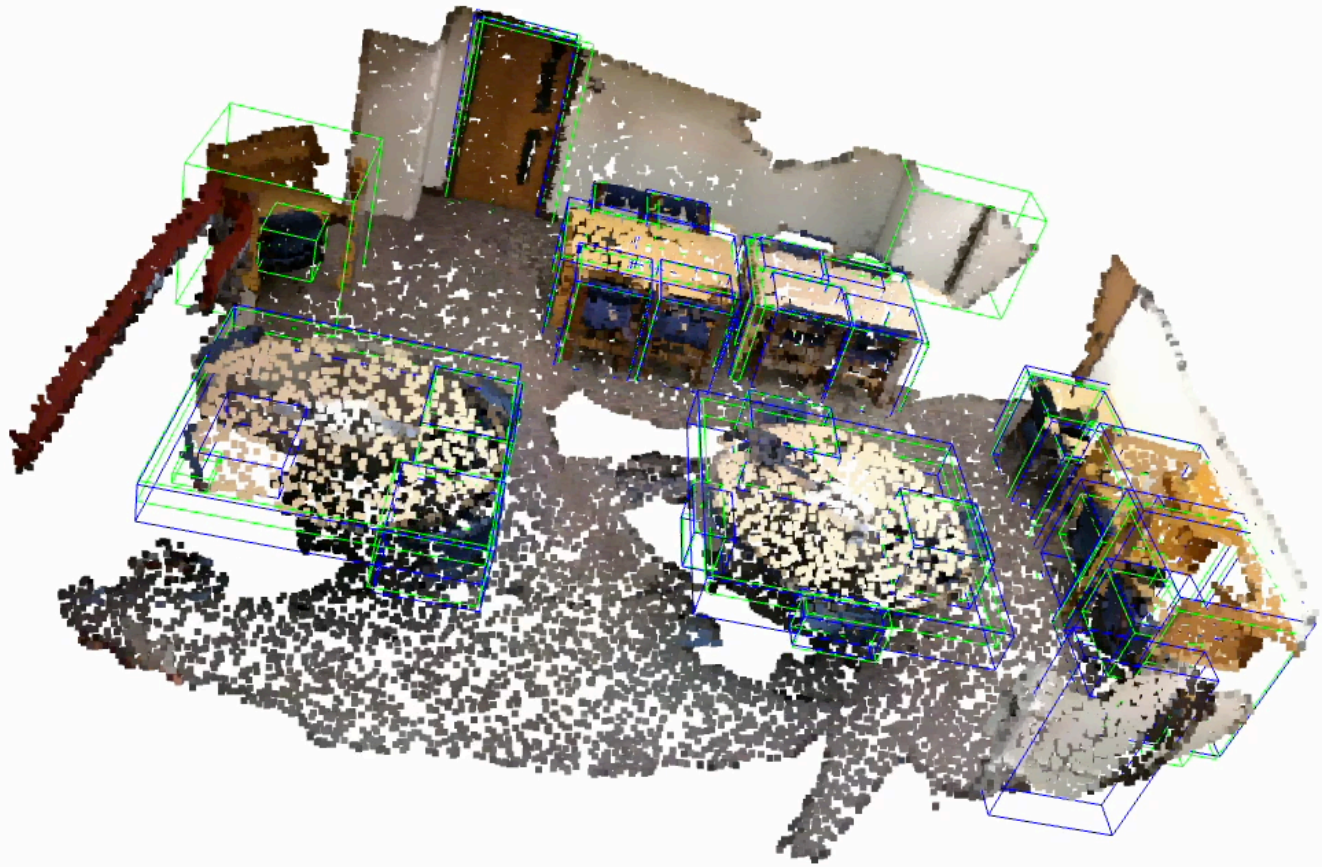
MS-A WITH DIFFERENT FEATURE SCALES.

Feature scale $= s$ means the feature map contains $s \times N$ points. A larger s denotes a feature map with higher point density (*i.e.*, resolution)

Scales s	mAP@0.25	mAP@0.50	mAP_S	mAP_M	mAP_L
[1]	68.6	51.8	63.1	76.6	83.2
[1, 2]	**68.9**	**52.5**	**65.0**	**77.5**	**83.9**
[0.5, 1, 2]	67.9	51.7	64.6	76.7	**83.9**







Thank you!