# BIP3D: Bridging 2D Images and 3D Perception for Embodied Intelligence

Supplementary Material

## A. Ablation on 3D Position Embedding

Position embedding in the spatial enhancer is a crucial component of BIP3D, serving to bridge the gap between 2D image features and 3D space. Table A.1 demonstrates the impact of 3D PE on detection performance. When the 3D PE is removed, the overall AP decreases by 3.09%.

| 3D PE | Overall | Head | Common | Tail |
|---|---|---|---|---|
| | 17.82 | 23.63 | 14.34 | 15.39 |
| ✓ | **20.91** | **27.57** | **18.77** | **16.03** |

Table A.1. Ablation Results of 3D PE.

To more intuitively illustrate that the spatial enhancer achieves spatial modeling, we visualize the correlations between 3D position embeddings **IPE**. As shown in Figure A.4, it can be observed that embedding correlations exhibit a positive relationship with their 3D positions.

## B. Inference Efficiency

We compared the inference speeds of BIP3D and EmbodiedScan on a 4090 GPU, as shown in Figure A.1. When considering the point cloud preprocessing time, BIP3D consistently exhibits lower latency than EmbodiedScan, with a more pronounced advantage when the number of views is small. When focusing solely on the neural networks, BIP3D's inference speed is slower at a higher number of views; however, when the number of views is reduced to below 8, BIP3D still maintains an efficiency advantage.

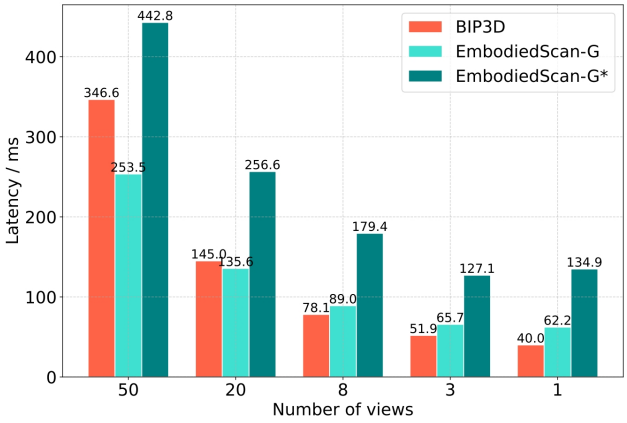



Figure A.1. Latency Comparison, where '*' indicates the inclusion of point cloud preprocessing time, encompassing multi-view aggregation and down-sampling.

## C. Scale-up

To test the impact of increasing model parameters on perception performance, we replaced the backbone with Swin-Transformer-Base. As shown in Table A.2, compared to Swin-Tiny, the overall AP improved by 1.21%, with a significant increase of 2.74% in long-tail categories. It is worth noting that GroundingDINO-Base showed an improvement from 58.1% to 59.7% over GroundingDINO-Tiny on COCO benckmark. To further enhance model performance, we incorporated additional training data from the ARKitScene dataset. This resulted in an additional 1.47% improvement. These results highlight the positive impact of both scaling up the model size and enriching the training dataset on improving detection accuracy.

| Backbone | ARKit | Overall | Head | Common | Tail |
|---|---|---|---|---|---|
| swin-tiny | | 20.91 | 27.57 | 18.77 | 16.03 |
| swin-base | | 22.12 | 28.63 | 18.77 | 18.77 |
| swin-base | ✓ | **23.59** | **30.20** | **19.59** | **20.88** |

Table A.2. Results of Scale-up Experiments.

## D. Detail of Camera Intrinsic Standardization

The parameters of standardized intrinsic are derived from the mean of the training set. Given that we use undistorted pinhole cameras, the parameters include $[focal_u, focal_v, center_u, center_v]$, which are set to $[432.579, 539.857, 256, 256]$. Intrinsic standardization may introduce issues such as pixel loss and zero padding, as shown in Figure A.2.

## E. Model Ensemble

For the model ensemble experiment listed in Table 3 of the main text, we employ five models. Two of these models are trained on the entire dataset, utilizing permutation corner distance loss and Wasserstein distance loss, respectively. The remaining three models are trained on distinct data subsets: ScanNet, 3RScan, and Matterport3D. The strategy for model ensemble is 3D NMS with 0.4 IoU threshold.

## F. Permutation Corner Distance Loss

For a single 3D bounding box, there are 48 possible permutations of its 8 corner points, denoted as $\mathcal{A}$, as shown in Figure A.3. Different permutations correspond to different $[w, l, h, roll, pitch, yaw]$ values. Therefore, using

Figure A.2. Images Comparison Before and After Camera Intrinsic Standardization. Left: Original, Right: Standardized.

$||\mathbf{B}_{pred} - \mathbf{B}_{gt}||$ directly as the loss function would result in incorrect gradients. We propose a permutation corner loss defined as:

$$L_{box} = \min_{1 \le i \le 48} [||\mathcal{A}(pred)_1 - \mathcal{A}(gt)_i||_2]$$

## G. Model Prediction Visualization

Figure A.5 visualizes the 3D detection results of the model, demonstrating that BIP3D can effectively handle a variety of complex indoor scenarios. Even for some objects that are unannotated, BIP3D is capable of detection, which provides feasibility for enhancing model performance through the use of semi-supervised learning in the future. Figure A.6 visualizes the 3D grounding results, illustrating the model's capability to identify and locate the specific target designated by the text among multiple objects of the same class.

## H. Algorithm Setting

Table A.3 lists more detailed model configurations and training parameters. Additionally, we employ two types of data augmentation during training: 1) applying a random grid mask to the depth map, and 2) performing random cropping on both the images and depth maps.

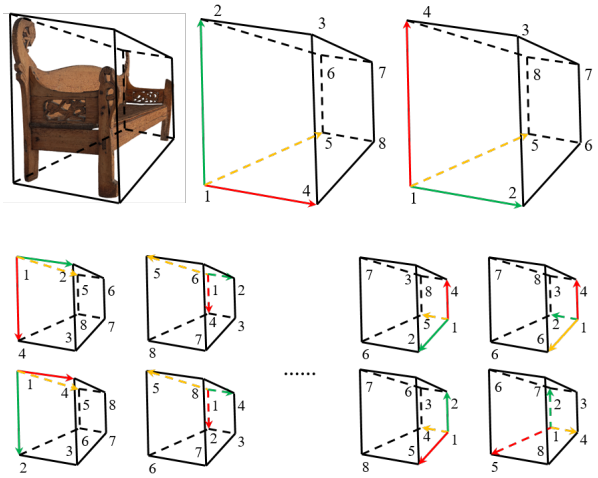
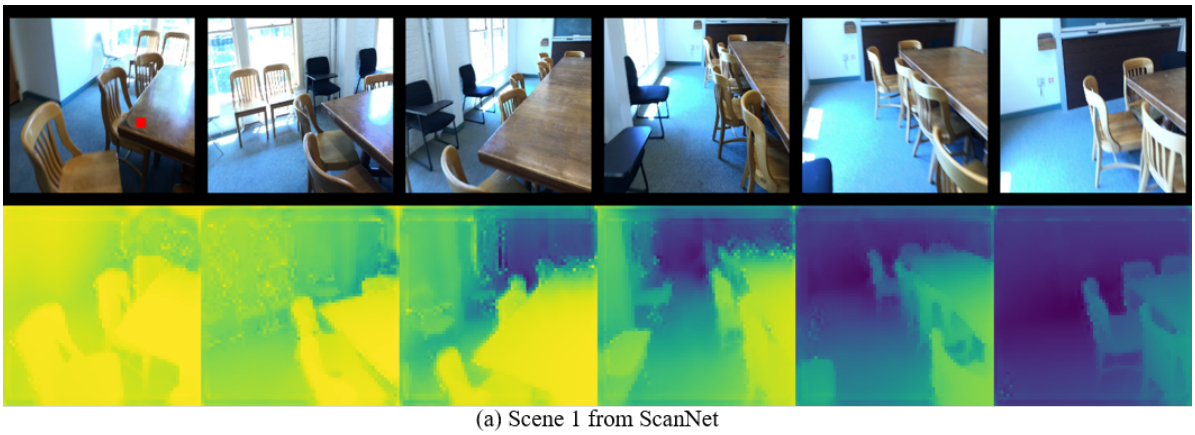


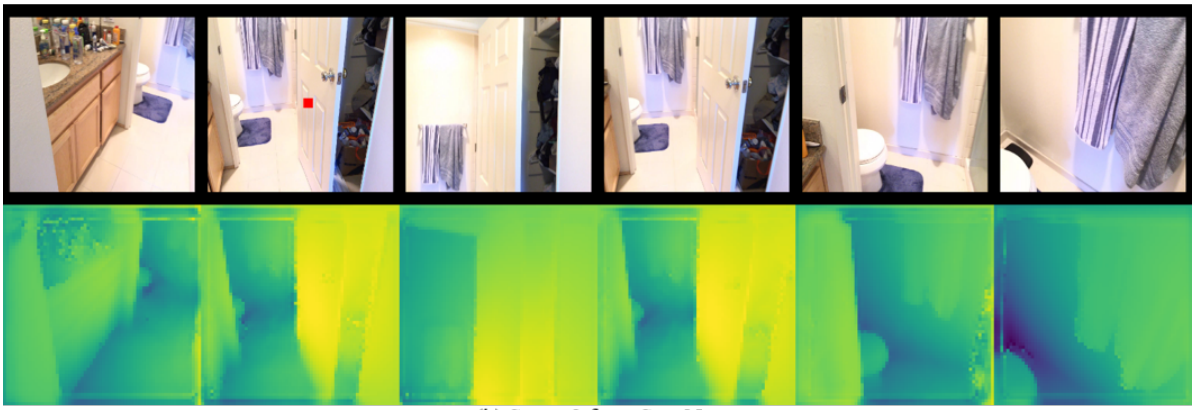
Figure A.3. The 3D Bounding Box Corners Permutations. For the same bounding box, there are a total of 48 different corner point permutation; the corner point order is indicated by numbers, with red, yellow, and green representing width, length, and height, respectively.

| Config | Setting |
|---|---|
| image backbone | swin-transformer-tiny |
| image neck | channel mapper |
| depth backbone | mini-ResNet34 |
| depth neck | channel mapper |
| text encoder | BERT-base |
| embed dims | 256 |
| feature levels | 4 |
| key points | 7 fixed and 9 learnable |
| feat enhancer layers | 6 |
| decoder layers | 6 |
| anchor per view | 50 |
| max depth $D$ | 10 |
| num of points $K$ | 64 |
| optimizer | AdamW |
| base lr | 2e-4 |
| image backbone lr | 2e-5 |
| text encoder lr | 1e-5 |
| detection epochs | 24 |
| grounding epochs | 2 |
| batch size | 8 |
| weight decay | 5e-4 |
| drop path rate | 0.2 |
| $\lambda_1$ | 1.0 |
| $\lambda_2$ | 0.8 |
| $\lambda_3$ | 1.0 |
| dn queries | 100 |
| training views | 18 |
| test views | 50 |

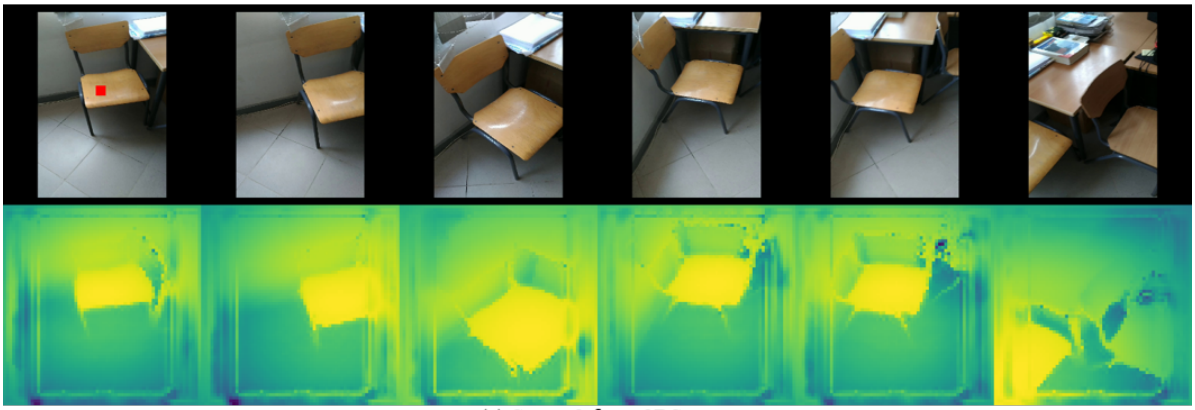
Table A.3. Model Configurations and Training Parameters.



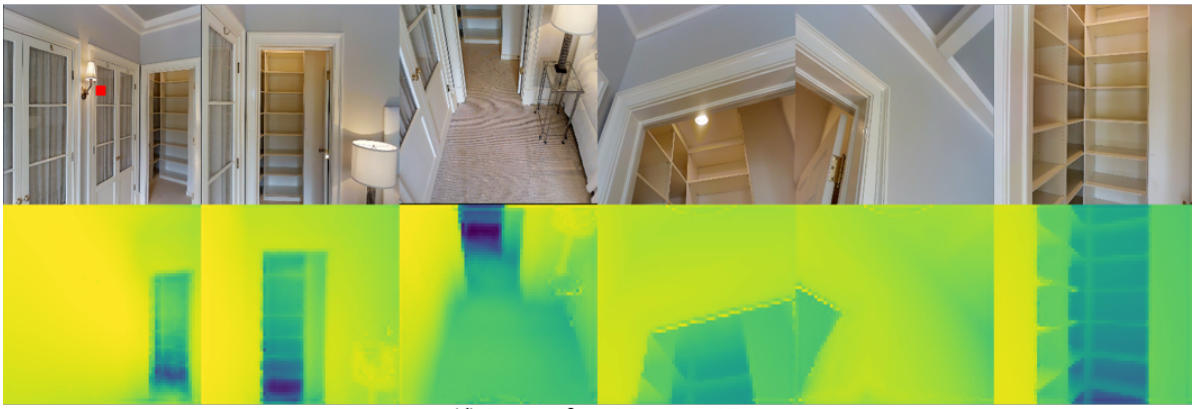
(a) Scene 1 from ScanNet



(b) Scene 2 from ScanNet

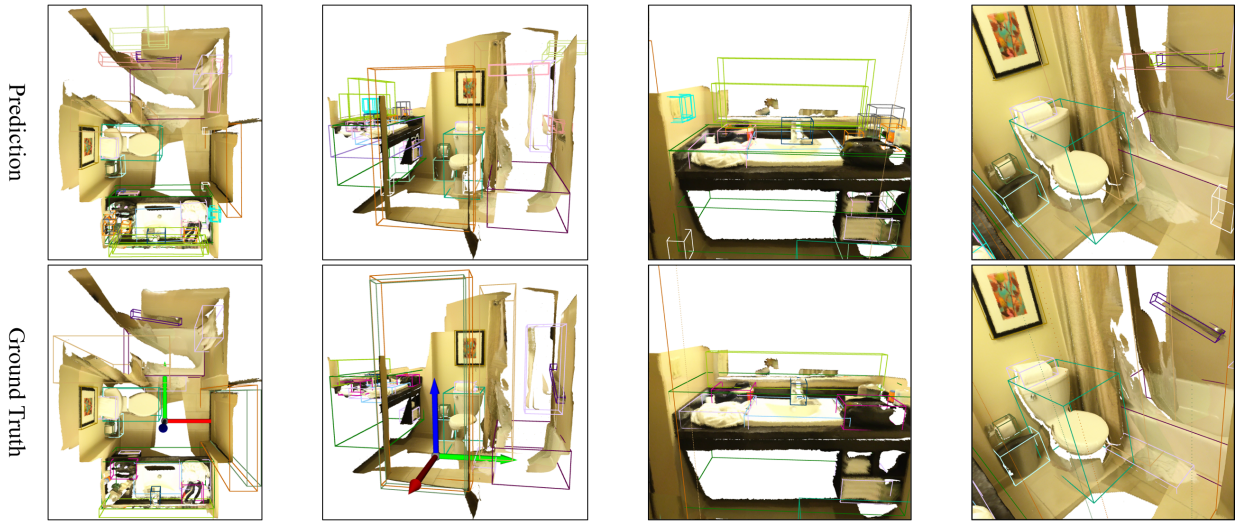


(c) Scene 3 from 3RScan

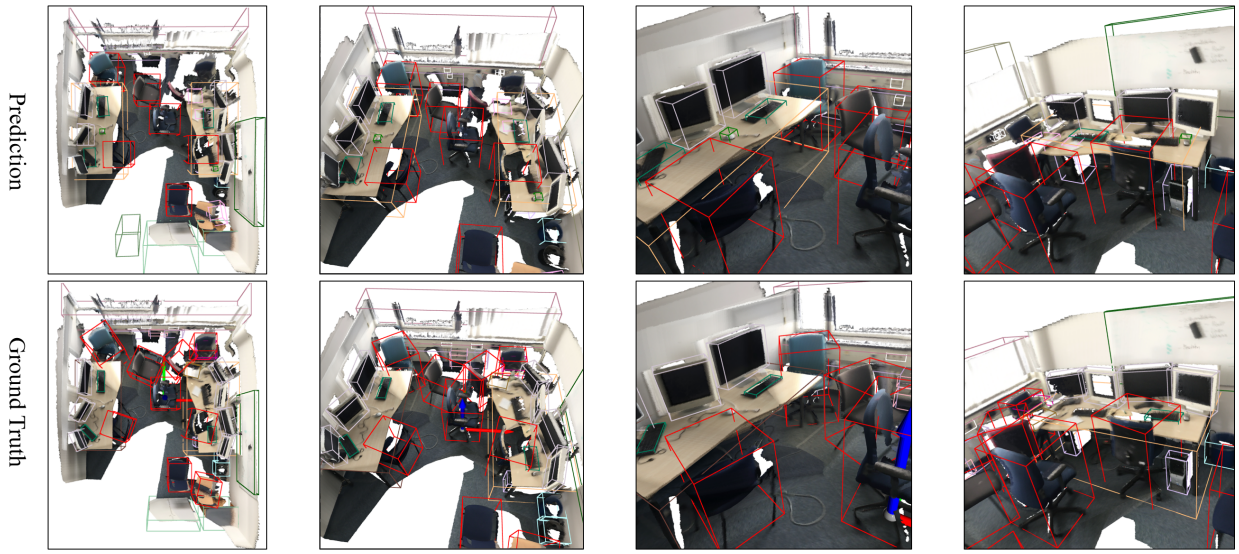


(d) Scene 4 from Matterport3D

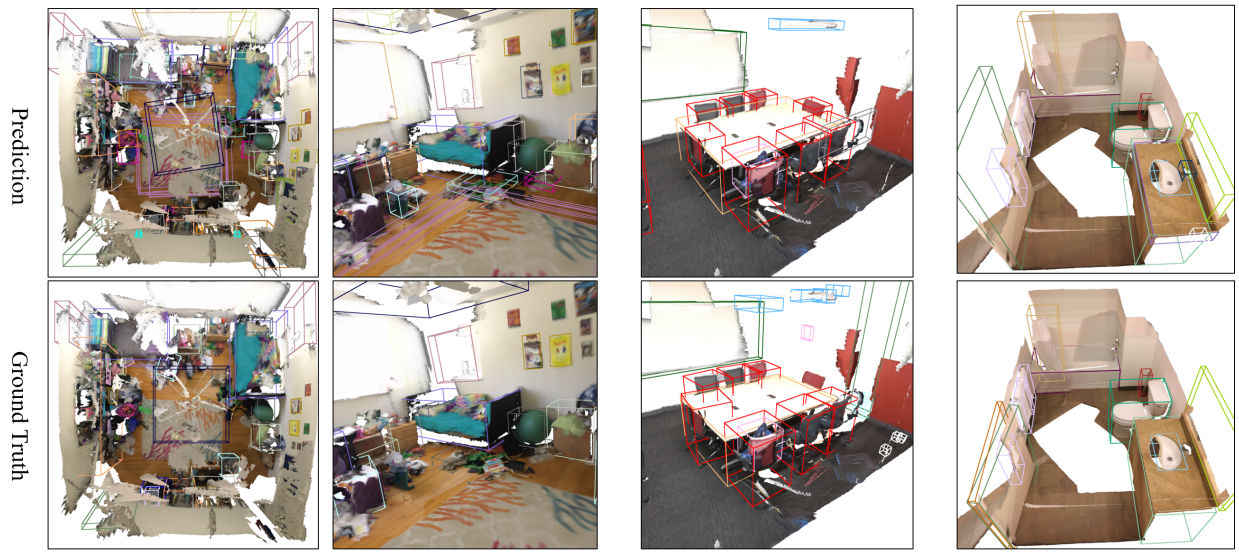
Figure A.4. Visualization of the Correlations of Position Embeddings. The red boxes on the images indicate the selected target location, while the heatmaps represent the cosine similarity between all position embeddings and the position embedding of the target location.




(a) Scene 1



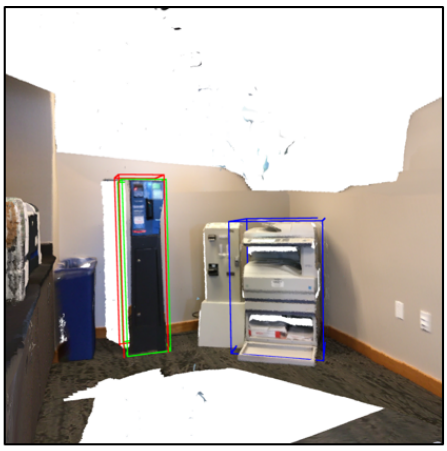

(b) Scene 2



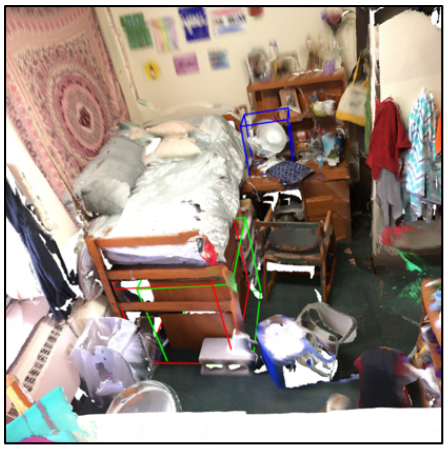

(c) Scene 3 (d) Scene 4 (e) Scene 5

Figure A.5. Visualization of 3D Detection Results. The color of the boxes indicates the category.

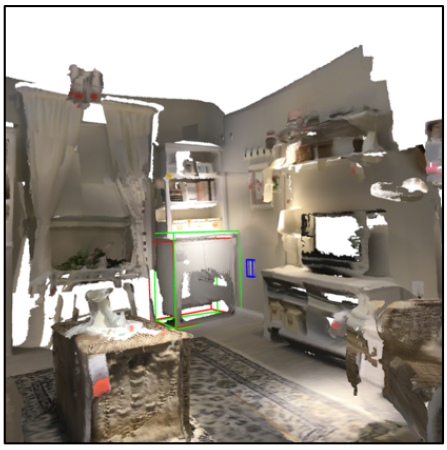
(a) find the bag that is closer to the bathtub



(b) select the machine that is farthest from the copier



(c) choose the dresser that is closer to the fan



(d) select the cabinet that is near the socket

(e) select the chair that is closer the crate

(f) find the book that is closer to the ladder . it is in the middle of the door and the jar . it is next to the jar

Figure A.6. Visualization of 3D Visual Grounding. Green boxes represent the ground truth, red boxes represent the predictions, and blue boxes represent reference objects, such as the 'bathtub' in 'find the bag that is closer to the bathtub'.