Appendix

A. Supplementary Implementation Details

In this section, we first introduce the descriptions of dataset we verified the effects of our pre-training scheme on. Then, we elaborate on the details of the baseline models we used in our empirical studies. Finally, we detail the specific configurations used in all of our experiments.

A.1. Datasets and metric

SUN RGB-D [51] is a challenging large-scale 3D indoor dataset, consisting of 10,335 RGB-D images with labeled 3D bounding boxes for 37 categories. Depth images are converted to point clouds using provided camera poses, and we follow the standard 5285, 5050 splits for the training and testing stages, respectively. We report the accuracy on the test set of SUN RGB-D using the mean Average Precision at two different IoU thresholds, 0.25 and 0.5, respectively.

ScanNetV2 [9] is a 3D interior scene dataset with rich annotations, consisting of 1513 indoor scenes and 18 object classes. The labels include semantic labels, per-point instances, 2D and 3D bounding boxes. For 3D object detection, we use the common metrics for evaluation [43], measuring the mean Average Precision (mAP) under two IoU thresholds of 0.25 and 0.5. For 2D detection, we follow [5] to report the Average Precision (AP) under 0.5, 0.75, and 1.0.

KITTI [16] is a widely adopted outdoor 3D object detection benchmark, consisting of 7481 training images and 7518 test images. To evaluate our approach, we follow [66] and adopt the Average Precision (AP) of 3D bounding boxes under three-level difficulties, easy, moderate and hard. Detection scores for the car category under the intersection-over-union (IoU) threshold of 0.7 are reported.

CIFAR-FS [4], FC100 [40], miniImageNet [55], are widely used few-shot image classification datasets, and we use these datasets to evaluate PiMAE's image feature extractor. The top-1 accuracy under 5-way 1-shot and 5-way 5-shot settings on different datasets is adopted as the evaluation metric.

A.2. Baseline Approaches

We evaluate our interactive multi-modal training pre-training scheme by fine-tuning three state-of-the-art 3D object detectors, and our baseline implementations rigorously follow their publicly released codes.

3DETR [39] is a simple, end-to-end 3D detection pipeline that does not require finely crafted 3D detection backbone. Instead, its versatile attention-based backbone maximally preserves the vanilla Transformer blocks to reach comparable performance with CNN-based detectors.

Group-Free 3D [35] is another approach implementing the Transformer models on 3D object detection task, using both a well-designed query locations for objects and an ensembling of detection results. Unlike PointNet-based networks [44, 45, 49] that create a local grouping scheme for each object candidate, Group-Free uses an attention mechanism on all the point cloud points.

Config	Value
optimizer	AdamW [28]
base lr	1e-3
weight decay	0.05
batch size	256
lr schedule	cosine decay
warmup epochs	15
epoch	400
augmentation	None

Table 8. **Pre-training configuration.**

DETR [5] is an end-to-end 2D object detection that uses a Transformer architecture to force unique predictions with bipartite matching. DETR can quickly and efficiently make predictions of the relations between objects and the global image context from a small set of object queries.

MonoDETR [66] is a novel, state-of-the-art DETR-based model for monocular 3D detection that does not rely on depth supervision, anchors, or Non-Maximum Suppression (NMS). It modifies DETR's vanilla transformer to incorporate depth estimates and predicts 3D annotations from the depth information inherent in images.

A.3. Pre-training Details

The encoder and decoder architectures in PiMAE follow the standard ViT [10] design, which consists of several Transformer blocks. In our PiMAE, the number of Transformer blocks for specific encoders and shared-decoders is both set to 3, while the numbers for specific decoders and shared-decoders are set to 2 and 1, respectively. For encoders, each Transformer block has 256 hidden dimensions and 4 heads for the multi-head self-attention module. For decoders, the numbers are adjusted to 192 and 3.

For the point cloud branch, we sample 2048 points from each 3D scene in SUN RGB-D [51], following previous work [64]. For the image branch, we adjust the resolution of each image to 256×352, and we use a patch size of 16 to patchify images. The specific configuration for pre-training PiMAE is given in Tab. 8.

A.4. Fine-tuning Protocol on SUN RGB-D, ScanNetV2, and KITTI

For fine-tuning on GroupFree3D [35], we insert our 3D feature extractor into the pipeline. TODO:inserted where? Compared to the original configuration, the only modification here is tuning the learning rate on the encoder lower to $lr = 3e-5$ to preserve the pre-trained prior. For detection on ScanNetV2, we lower the encoder learning rate to $lr = 6e-5$.

For experiments with 3DETR [39], our encoder consists of six Transformer blocks pre-trained with PiMAE. The exact setups as the original [39] are then used for fine-tuning, except that we apply a reduced learning rate of $lr = 1e-5$ to the encoder.

For DETR [5], we replace the vanilla Transformer encoder with our image branch feature extractor and our joint-encoder pre-trained on SUN RGB-D. The depth of the encoder is unchanged and we only apply a reduced learning rate of $lr = 1e-5$ to the encoder. We perform 2D object detection on the ScannetV2 2D detection dataset.

For MonoDETR [66], we replace its depth encoder with our 3-layer 3D feature extractor and follow the original configuration for training.

B. Additional Ablation Study

In this section, we give more ablation studies for further analysis of PiMAE.

Ablation study on Pipeline Architecture. During the reconstruction stage, as proposed in Sec. 3.4, a shared decoder architecture is adopted. The encoded features are first disentangled by the cross-modality decoder, and reconstructions are completed afterward with task-specific decoders. From Tab. 9, we find the additional shared-decoder design performance-enhancing, as it considers the cross-modal influence of masked tokens. Specifically, shared-decoder is a novel contribution of PiMAE, and we find it non-trivial, because the interactions in the masked tokens improve feature extraction.

Ablation Study on Masking Ratio. As reported in Tab. 10, we examined several masking ratios for PiMAE and find that the the model learns the best latent features when the masking ratio is set to 60%.

Encoder	Decoder	AP_{25}	AP_{50}
3+3	0+3	58.0	30.2
3+3	1+2	**59.4**	**33.2**
3+3	1+3	58.1	32.8
2+2	1+2	57.5	30.8

Table 9. **Ablation studies on model architecture.** a+b in encoder denotes specific encoders of a-layers ViT, and shared-encoder of b-layers ViT. c+d in decoder denotes c-layers ViT for shared-decoder and d-layers ViT for specific decoders. Experiments are based on 3DETR and performed on SUN RGB-D.

Mask Ratio	AP_{25}	AP_{50}
50%	58.7	33.1
60%	**59.4**	**33.2**
70%	58.4	33.0
80%	57.5	32.4

Table 10. **Ablation study on masking ratios.** Experiments with different masking ratio are conducted, and we report detection accuracy based of 3DETR on SUN RGB-D val set.

C. Additional Visualization

Reconstruction Results. In Fig. 8. We provide more examples of reconstruction visualizations. PiMAE simultaneously reconstructs masked point clouds and images with clear reconstructions reflecting semantic understanding.

Activation of Feature Map. This section provides more attention map examples generated by PiMAE's shared-encoder, where features from the two modalities first interact explicitly. By examining the self-attention weights, we can gain better insights on PiMAE's multi-modal interactions. We compute the self-attention from a point cloud token to all image tokens and show the attention values. In Fig. 9, PiMAE is able to attend to more foreground objects and with higher attention values, while other designs either attend to unrelated backgrounds(e.g. row 3, col 4), or have rather low attention values (e.g. row 2, col 4).

We also compute the self-attention from a image token to all point cloud tokens and display the attention weights. As shown in Fig. 10, given a image token as query, PiMAE accurately attends to the corresponding objects in the point cloud with highest values, showing a strong understanding of both 2D and 3D features.

Object Detection. We display more qualitative results comparing PiMAE and baseline in Fig. 7. On top of 3DETR [39], with PiMAE pre-training, we are able to detect more objects with more precise boxes.

Methods	bed	table	sofa	chair	toilet	desk	dresser	nightstd	bookshf	bathtub	AP_{25}	AP_{50}
DSS [52]	78.8	50.3	53.5	61.2	78.9	20.5	6.4	15.4	11.9	44.2	42.1	-
2D-driven [30]	64.5	37.0	50.4	48.3	80.4	27.9	25.9	41.9	31.4	43.5	45.1	-
PointFusion [61]	68.6	31.0	53.8	55.1	83.8	17.2	23.9	32.3	**37.7**	37.3	45.4	-
F-PointNet [44]	81.1	51.1	61.1	64.2	90.9	24.7	32.0	58.1	33.3	43.3	54.0	-
VoteNet [43]	83.0	47.3	64.0	75.3	90.1	22.0	29.8	62.2	28.8	74.4	57.7	-
3DETR [39]	81.8	50.0	58.3	68.0	90.3	28.7	28.6	56.6	27.5	77.6	58.0	30.3
+Ours	85.4	48.9	62.5	69.0	93.8	28.2	33.0	62.8	30.4	80.3	59.4(+1.4)	33.2(+2.9)
GroupFree3D [35]	**87.8**	53.8	70.0	**79.4**	91.1	**32.6**	36.0	66.7	32.5	80.0	63.0	45.2
+Ours	85.4	**55.1**	**73.3**	78.1	**96.0**	31.5	**40.8**	**67.8**	28.4	**89.1**	**64.6**(+1.6)	**46.2**(+1.0)

Table 11. **3D objection detection results on SUN RGB-D validation set.** Single-class precision is reported with AP_{25}. Results of previous methods are taken from [35, 39, 43].

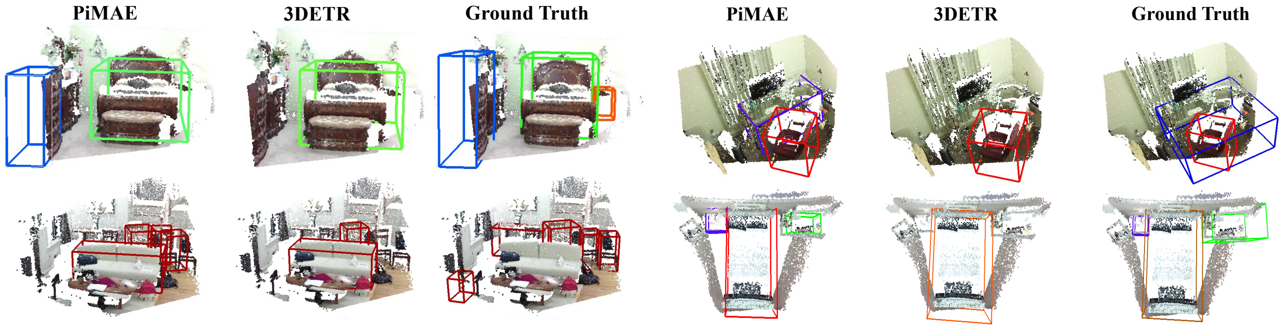

Figure 7. **Visualization of predictions on SUN RGB-D validation set.** We correctly detect more objects.

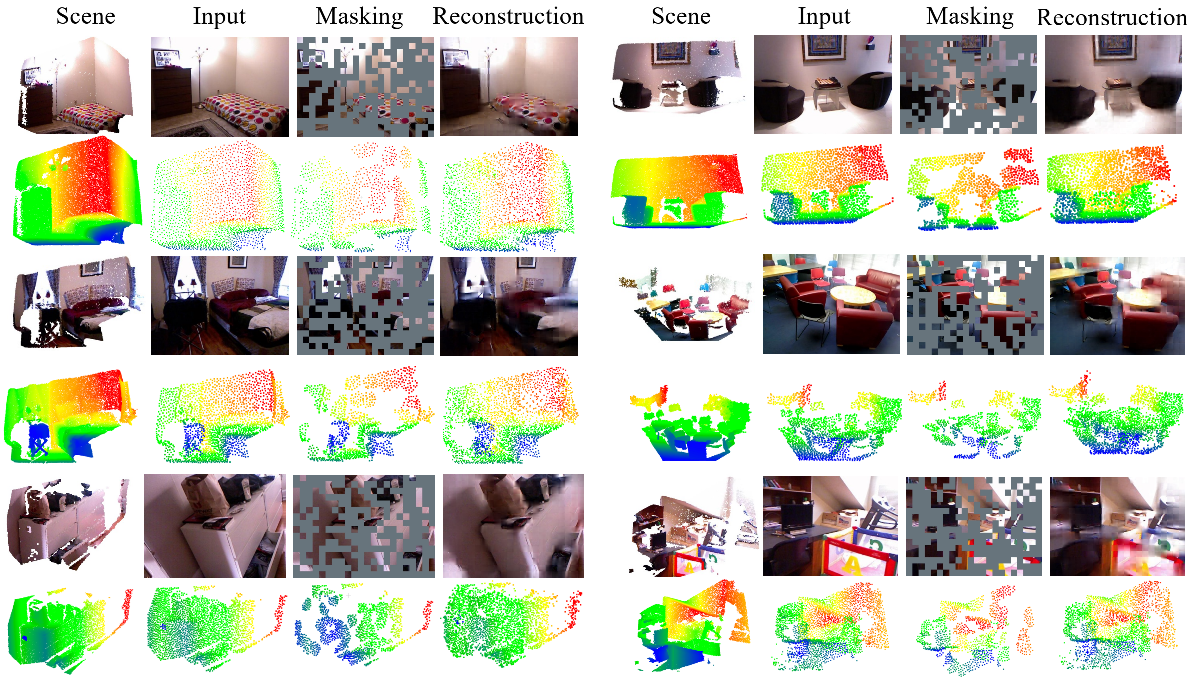

Figure 8. **Visualization of reconstruction results.** Our model is trained with 60% masking ratio. Point clouds are colored for better visualization purpose. PiMAE generalizes well for different scenes and reconstructs masked images (odd rows) and masked point clouds (even rows) simultaneously.

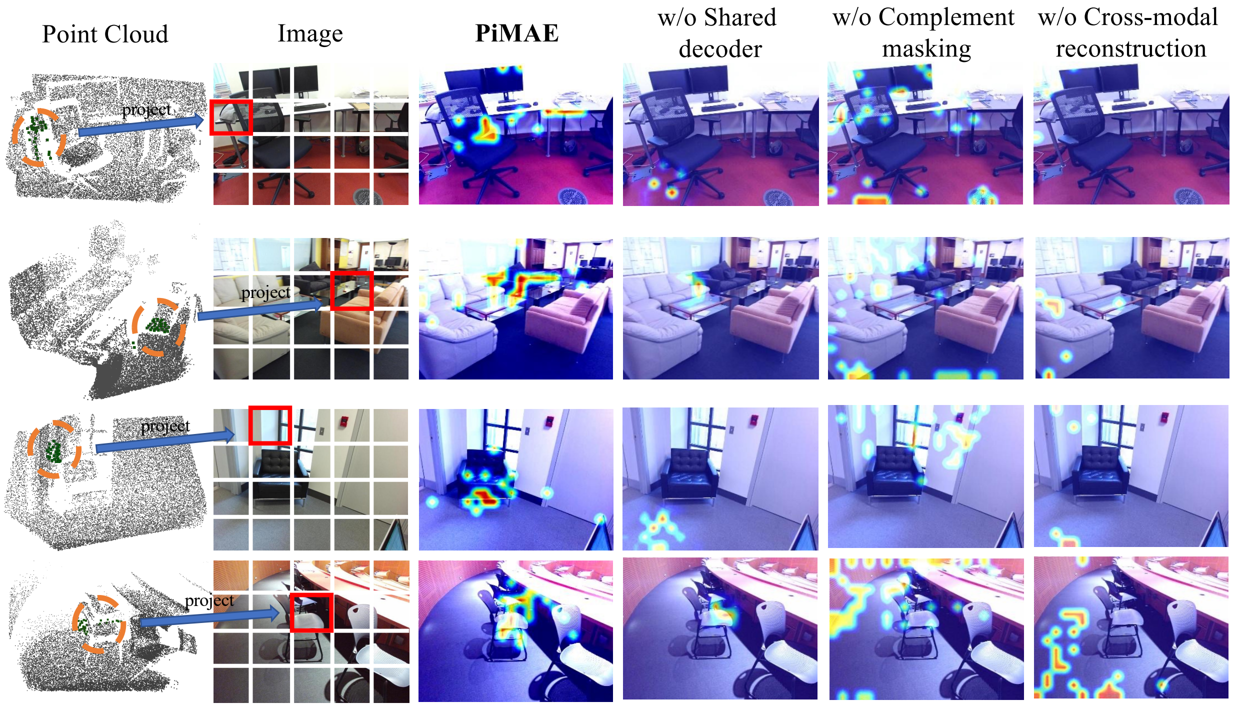

Figure 9. **Visualization of encoder attention, point cloud as query.** The encoder's attention between two modalities is visualized by computing self-attention from the query points (orange circle) to all the visible image tokens. Highest values are shown in red. We show the corresponding location (red square) of the query points after projection. From left to right shows ablation of PiMAE with different designs, including our final proposal, and settings that exclude shared-decoder, complement masking strategy and cross-modal reconstruction, respectively.

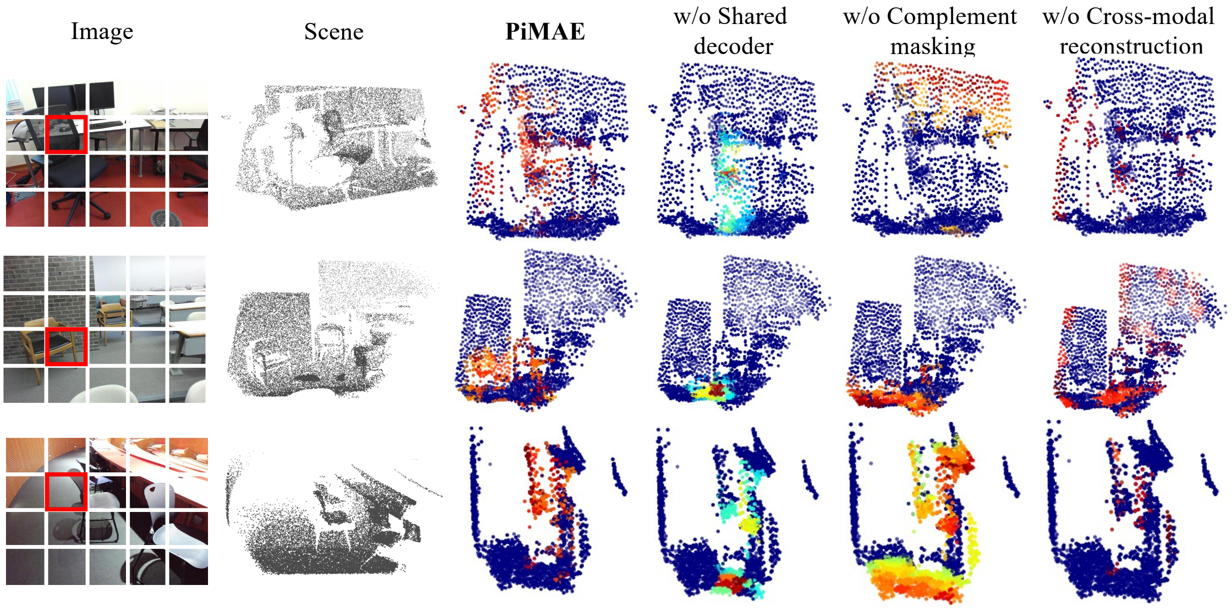

Figure 10. **Visualization of encoder attention, image as query.** The encoder attention between the two modalities is visualized by computing self-attention from the query of an image token (red square) to all the point cloud tokens. Highest values are shown in red. The attention intensity in the point cloud corresponds with the image patch query, showing the effectiveness of our cross-modal interactions during pre-training.