RI-Mamba: Rotation-Invariant Mamba for Robust Text-to-Shape Retrieval

Supplementary Material

7. Training and Implementation Details

We implement our method in PyTorch [29] and conduct all experiments on a single NVIDIA RTX 4090 GPU with 24GB memory. Our encoder consists of 12 RI-Mamba blocks with a feature dimension of 512. For RI positional and orientational embeddings, we use a two-layer MLP with a hidden size of 128 and GELU activation [15]. The FiLM module includes a bottleneck reducing the feature dimension from 512 to 128, followed by a two-layer MLP (hidden size 128) to compute the modulation parameters γ and β.

For pretraining, we adopt the cross-modal contrastive learning framework introduced in TAMM, training the 3D encoders for 200 epochs with 10 warm-up epochs. We use the AdamW optimizer [23] and a cosine learning rate schedule with a base learning rate of $5\mathrm{e}{-4}$. The CLIP image adapter from TAMM is used to reduce the domain gap between rendered and real images. All experiments use the OpenCLIP-ViT-G/14 variant [6] as pre-trained image-text encoders. For patch-based encoders (PointBERT, Duo-Mamba, RI-Transformer, RI-Mamba), the input point cloud consists of 10K points. For the graph-based encoder LocoTrans, we use 1024 points following the setting in their paper. We follow original configurations and set the number of point patches (G) to 384 for PointBERT, 256 for Duo-Mamba, and 64 for RI-Mamba and RI-Transformer. The number of neighbors (k) is set to 64 for PointBERT, 20 for LocoTrans, and 32 for RI-Mamba and RI-Transformer.

Results of PointBERT pretrained on ShapeNet in Tab. 2-4 are obtained using the released weights from TAMM. For all other zero-shot results, we pretrain the encoders using the configuration described above. All results are reported from a single run with a fixed random seed of 0. For the supervised text-to-shape experiment (Tab. 1), RI-Mamba is trained with the same setup. To evaluate on Text2Shape-SO(3) (Tab. 1), we directly use the released weights of SCA3D, while Parts2Words is trained using the official implementation and default hyperparameters due to the lack of available pretrained weights.

8. OmniObject3D Captions

For each object in the OmniObject3D dataset [43], a detailed textual annotation is provided, including a high-level summary as well as descriptions of appearance, material, style, and function, as illustrated in examples 1 and 2 below. To construct text queries for zero-shot text-to-shape retrieval (Sec. 4.2), we concatenate the **Summary** and **Appearance** fields. This ensures that each query captures both semantic class information (from the summary) and fine-grained visual details (from the appearance description), enabling better discrimination among objects belonging to the same category.

OmniObject3D: Example 1

Summary: A dark green table.
Appearance: This is a rectangular table, made of a flatter rectangle for the tabletop, the rest of the rectangle is sealed, made of bamboo, with square edges for the skeleton, the overall color is dark green, left and right symmetrical.
Material: Bamboo, hard, slightly reflective, rough surface.
Style: Realistic.
Function: Placing items to assist with work.

Constructed Query

A dark green table. This is a rectangular table, made of a flatter rectangle for the tabletop, the rest of the rectangle is sealed, made of bamboo, with square edges for the skeleton, the overall color is dark green, left and right symmetrical.

OmniObject3D: Example 2

Summary: It's a toy train.
Appearance: This toy train as a whole is green, the top part is yellowish green, there is a round bump, the lower part of the front and back of each side of a cylindrical bump, this train has a total of four wheels, the train surface does not have any hand-painted patterns, the overall structure of the axisymmetric.
Material: Plastic, rubber, iron, hard, smooth surface, slightly reflective.
Style: Cartoon.
Function: Entertainment, decoration.

Constructed Query

It's a toy train. This toy train as a whole is green, the top part is yellowish green, there is a round bump, the lower part of the front and back of each side of a cylindrical bump, this train has a total of four wheels, the train surface does not have any hand-painted patterns, the overall structure of the axisymmetric.

9. Additional Visualization on Text2Shape

We provide extra visualization for text-to-shape retrieval results of SCA3D and RI-Mamba on the Text2Shape benchmark in Fig. 6, Fig. 7, and Fig. 8. As we can see, RI-Mamba demonstrates strong robustness to rotations, whereas SCA3D performs well only under canonical poses. Under rotation, SCA3D struggles to distinguish tables and chairs, often failing to retrieve the correct object described in the text. These results underscore the challenges of real-world retrieval scenarios, where objects appear in varied orientations, and highlight the practical effectiveness of our proposed RI-Mamba.

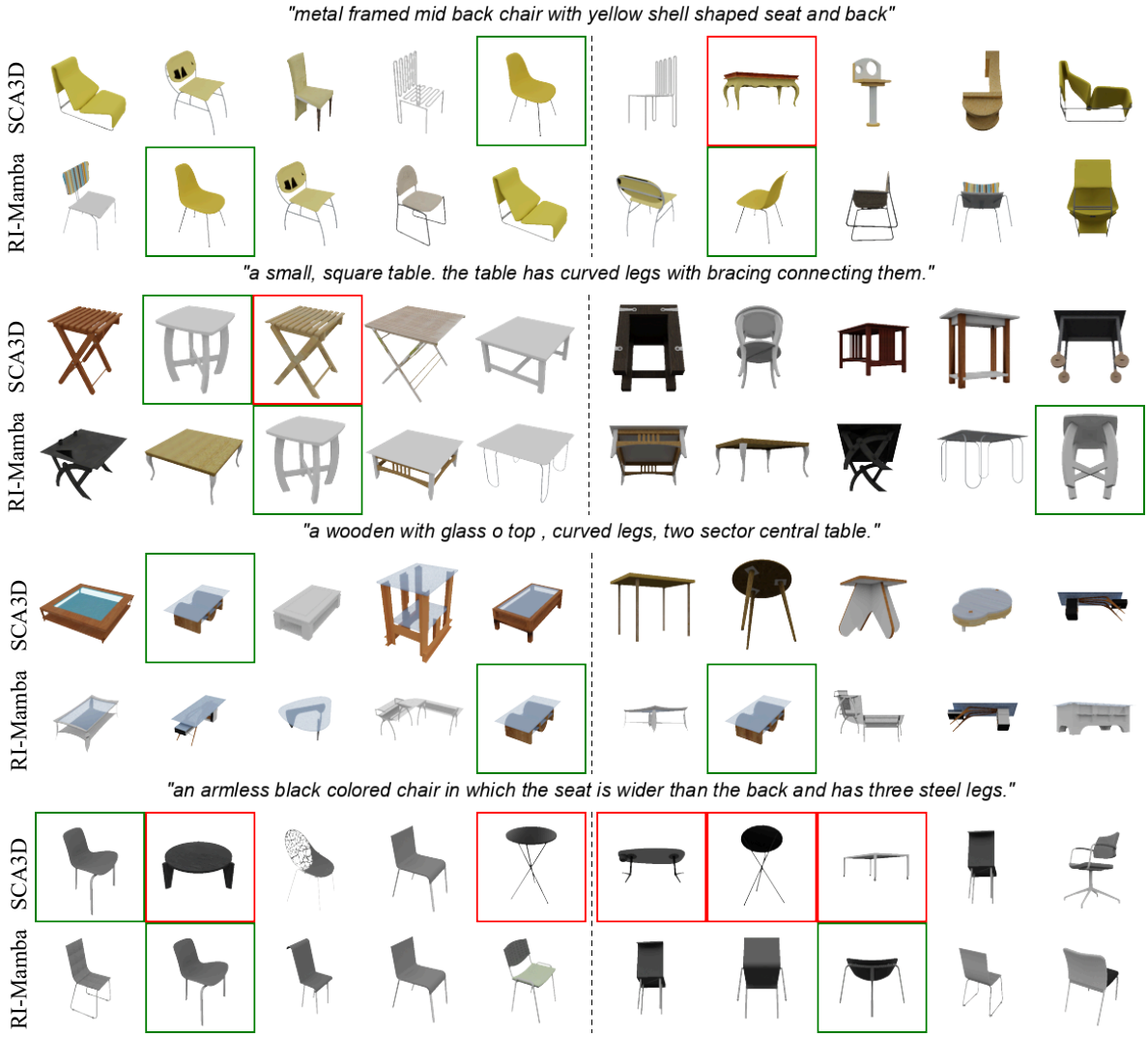

Figure 6. Retrieval results of SCA3D and RI-Mamba on the Text2Shape dataset. The target object is highlighted in the green box, and objects from the incorrect class are marked with red boxes.

Figure 7. Retrieval results of SCA3D and RI-Mamba on the Text2Shape dataset. The target object is highlighted in the green box, and objects from the incorrect class are marked with red boxes (continued).

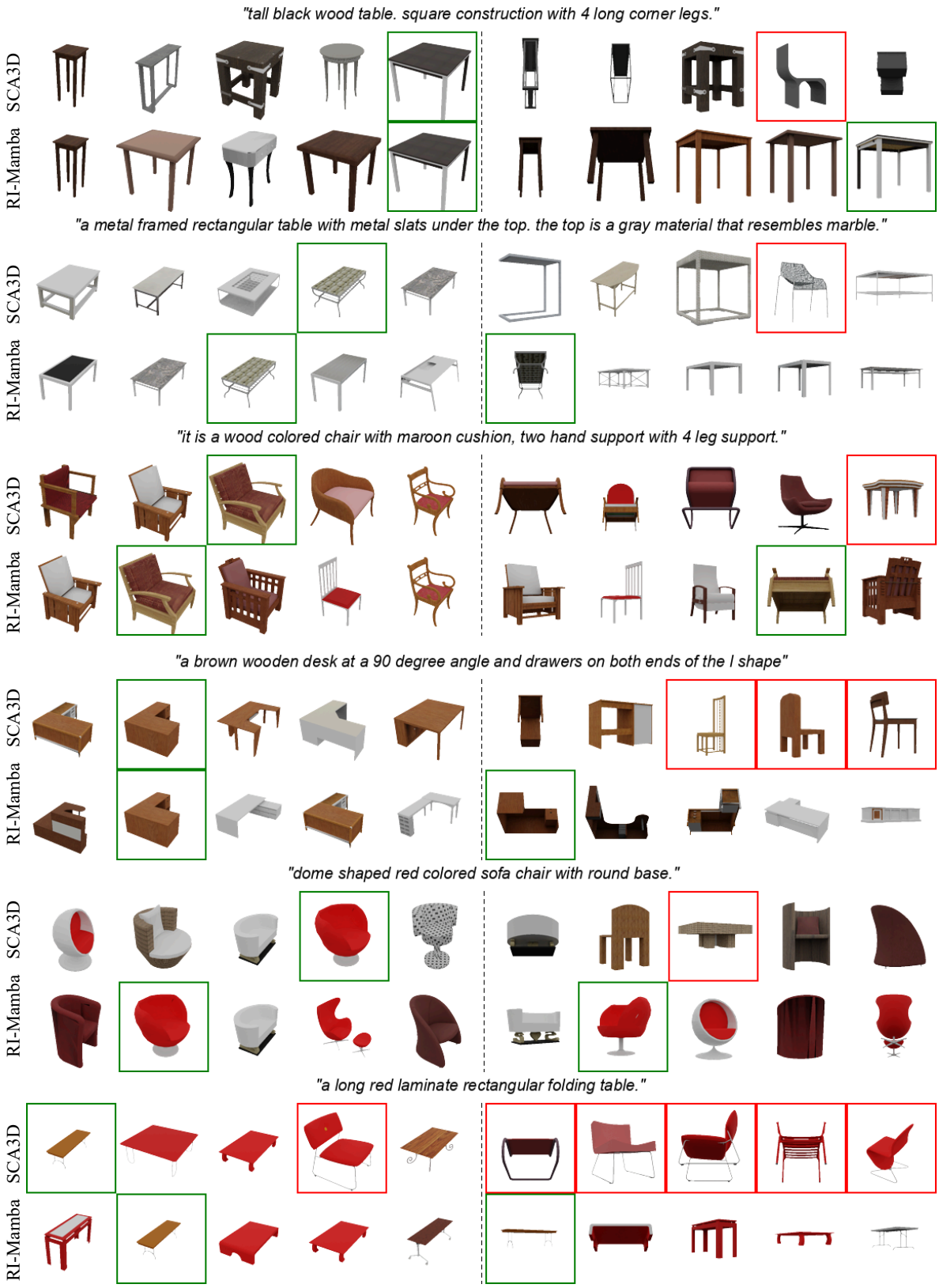

Figure 8. Retrieval results of SCA3D and RI-Mamba on the Text2Shape dataset. The target object is highlighted in the green box, and objects from the incorrect class are marked with red boxes (continued).