Supplemental Materials to "Dense Multimodal Alignment for Open-Vocabulary 3D Scene Understanding"

Ruihuang Li[1], Zhengqiang Zhang[1], Chenhang He[1], Zhiyuan Ma[1], Vishal M. Patel[2], and Lei Zhang[1(✉)]

[1] Hong Kong Polytechnic University
[2] Joins Hopkins University
{csrhli, cslzhang}@comp.polyu.edu.hk, vpatel36@jhu.edu
https://github.com/lslrh/DMA

In this supplemental file, we provide the following materials:

- The instructions to reduce noisy tags via GPT and the corresponding visualizations of 3D label map by using denoised tags (referring to "**Sec. 3.1**-Comprehensive Text Modality Generation-*Reliable GPT-based Denoising*" in the main paper);
- Examples of scene-level captions extracted by MLLMs and the corresponding visualizations of 3D label map by using the captions as queries (referring to "**Sec. 3.1**-Comprehensive Text Modality Generation-*Scalable Scene Descriptions*" in the main paper);
- The impacts of tagging models and MLLMs on different datasets (referring to "**Sec. 3.1**-Comprehensive Text Modality Generation" and "**Sec. 4.3**-Ablation Study-*tagging models* vs. *MLLMs*" in the main paper);
- Effect of Mutually Inclusive Loss (**MIL**) (referring to "**Sec. 3.4**-Dense Multimodal Alignment-*Mutually Inclusive Loss*" in the main paper);
- More qualitative results on indoor and outdoor datasets (referring to "**Sec. 4.2**-Comparison with State-of-The-Arts" in the main paper).

1 Reducing Noisy Tags

Instruction. We use GPT to reduce noisy tags generated by RAM [4]. The corresponding instructions and examples are given in Fig. 1. Given the input list of tags, we instruct GPT to evaluate each word individually to determine if it meets the given rules, and we ask it to output the chain of thought and the final boolean list.

Denoised tagging results. We give some examples of the denoised scene tags and the corresponding visualizations in Fig. 2. We can observe that many noisy tags can be filtered out in terms of the rules given in Fig. 1. We also visualize the text-to-3D label map M_{tag}^{3D} as defined in Eq. 2 of the main paper. Label errors in this 3D label map are consequently reduced based on the denoised tags.

Instruction: I will give you a list containing 10 words. Please output a boolean list containing 10 boolean values corresponding to the input list, according to the following rules:
(1). if the word is not noun, output False. (e.g., "wooden", "small", etc.)
(2). if the word represents colors, locations, room or space types, output False. (e.g., "green", "black", "room", "kitchen", "office", "bathroom", etc.)
(3). if the word indicates the object is animals or humans, output False. (e.g., "man", "woman", "cat", "dog", etc.)
(4). otherwise, output True.
Please examine the words in the list one by one and give the chains of thought, and output corresponding boolean list according to the above rules.

Examples:
User input: [room, desk, computer, chair, small, black, curtain, living room, hotel room, man]
Output:
1. "room": the word is noun, but "room" represents room or space types, therefore violates the rule (2), output "False".
2. "desk": the word is noun, and "desk" doesn't represent colors, locations, room or space types, and "desk" is not animals or humans, therefore according to rule (4), output "True".
3. "computer": the word is noun, and "computer" doesn't represent colors, locations, room or space types, and "computer" is not animals or humans, therefore according to rule (4), output "True".
4. "chair": the word is noun, and "chair" doesn't represent colors, locations, room or space types, and "chair" is not animals or humans, therefore according to rule (4), output "True".
5. "small": the word is adjectives but not noun, therefore violates the rule (1), output "False".
6. "black": the word is noun, but "black" represents colors, therefore violates the rule (2), output "False".
7. "curtain": the word is noun, and "curtain" doesn't represent colors, locations, room or space types, and "curtain" is not animals or humans, therefore according to rule (4), output "True".
'8. "living room": the word is noun, but "living room" represents room or space types, therefore violates the rule (2), output "False".
'9. "hotel room": the word is noun, but "living room" represents room or space types, therefore violates the rule (2), output "False".
'10. "man": the word is noun, and "man" doesn't represent colors, locations, room or space types, but "man" is animals or humans, therefore violates the rule (3), output "False".
'<results>: [False, True, True, True, False, False, True, False, False, False]'

Fig. 1: Instruction and examples to reduce noisy tags.

2 Scene Captions

In addition to category names, we also leverage MLLMss to generate comprehensive descriptions for each view. Fig. 3 shows some examples of the scene captions extracted by LLaVA [3] and the visualizations of 3D label map M_{llm}^{3D} generated by employing the captions as queries. We can build dense text(caption)-to-3D correspondences, thus achieving more accurate and fine-grained alignment.

3 Tagging Models *vs*. MLLMs

In Tab. 1 we compare the results of aligning 3D points to image tags and captions on different datasets. The former involves complete category information, while the latter contains comprehensive scene descriptions and more contextual information. As can be seen, tagging model plays a crucial role in the overall performance, as it encompasses more extensive categories and semantics. MLLM further enhances the final performance by incorporating rich contextual semantics and diversified descriptions. It is worth noting that MLLM only contributes to a marginal

performance improvement on the nuScenes [1] dataset since this dataset contains a limited number of image views.

	ScanNet	Matterport3D	nuScenes
Tagging+GPT	46.3	37.7	43.3
MLLM	24.2	19.4	11.3
Both	50.5	39.8	43.8

Table 1: Performance comparison of tagging model and MLLM.

4 Mutually Inclusive Loss

As discussed in Sec. 3.4 of the main paper, we use mutually inclusive loss to align 3D representations to text modality. In Tab. 2, we compare the performance by using mutually **inclusive** (BCE loss) and **exclusive** (CE loss) losses. Compared to binary cross-entropy loss, the cross-entropy loss leads to a performance degradation of 6.4% mIoU because it causes conflicts between text labels with similar semantics, making it difficult for the model to converge and limiting the performance improvement.

	Cross-entropy	Binary cross-entropy
mIoU	46.9	53.3

Table 2: Comparison of mutually inclusive and exclusive losses on ScanNet dataset.

5 Qualitative Results

We give more qualitative results on both indoor and outdoor datasets in Fig. 4 and Fig. 5. Our method consistently produces better results on ScanNet [2] and nuScenes [1] datasets, demonstrating the robustness and better generalization ability of the proposed method on different data distributions and open-set categories.

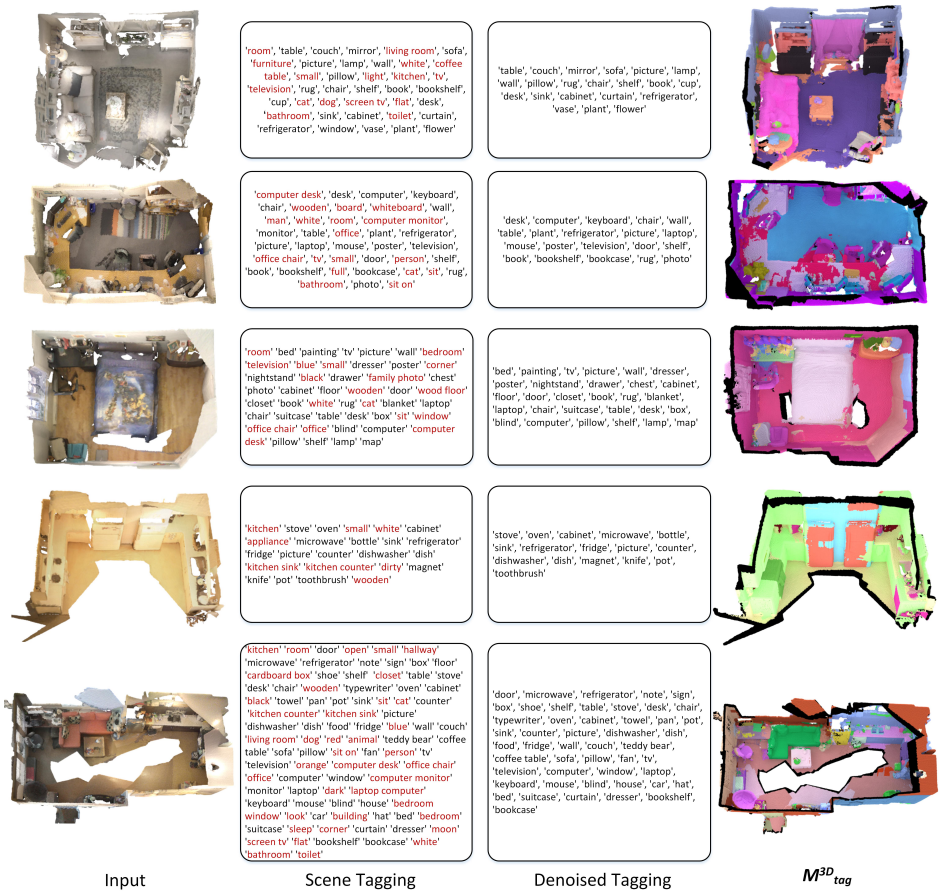

Fig. 2: Examples of scene tagging results by RAM [4] and visualizations of 3D label map M^{3D}_{tag} by employing the tags as queries. The red words denote the noisy tags which are filtered out by GPT and multi-view voting.

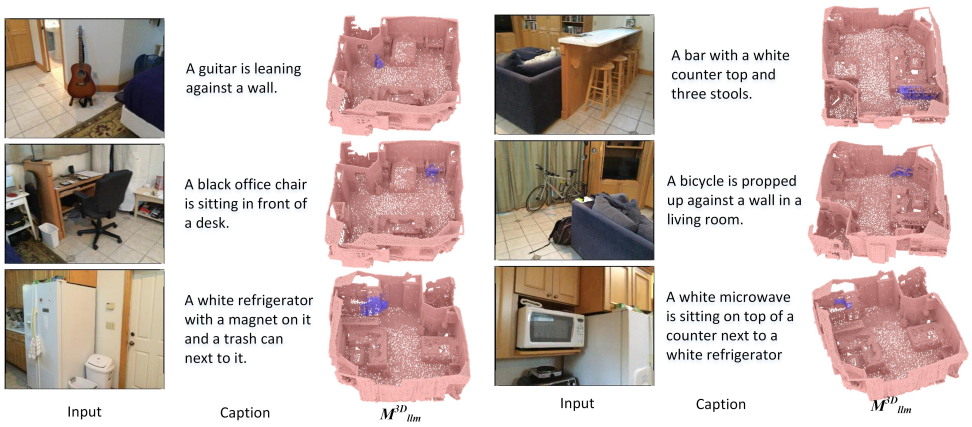

Fig. 3: Examples of scene captions extracted by LLaVA [3] and visualizations of 3D label map M^{3D}_{llm} by employing the captions as queries. We can build dense text-to-3D correspondences based on the obtained captions.

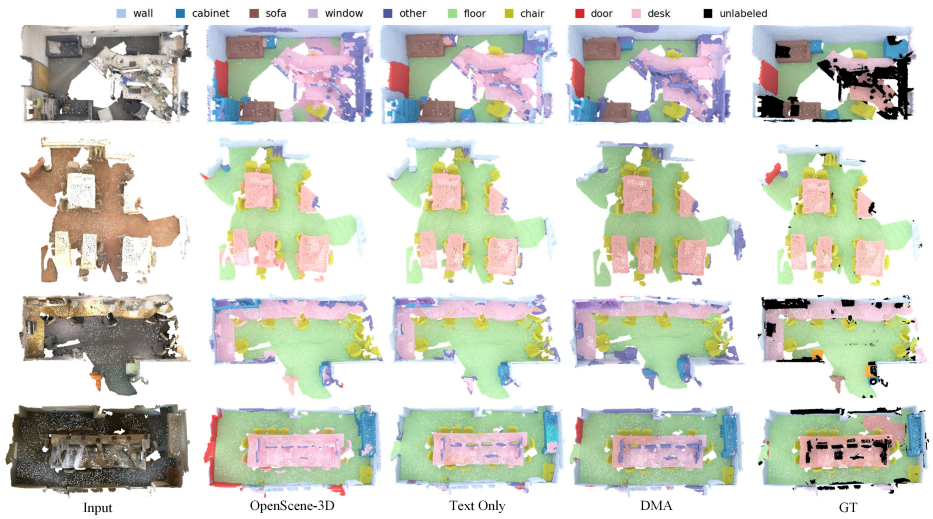

Fig. 4: Qualitative results of different methods on ScanNet dataset.

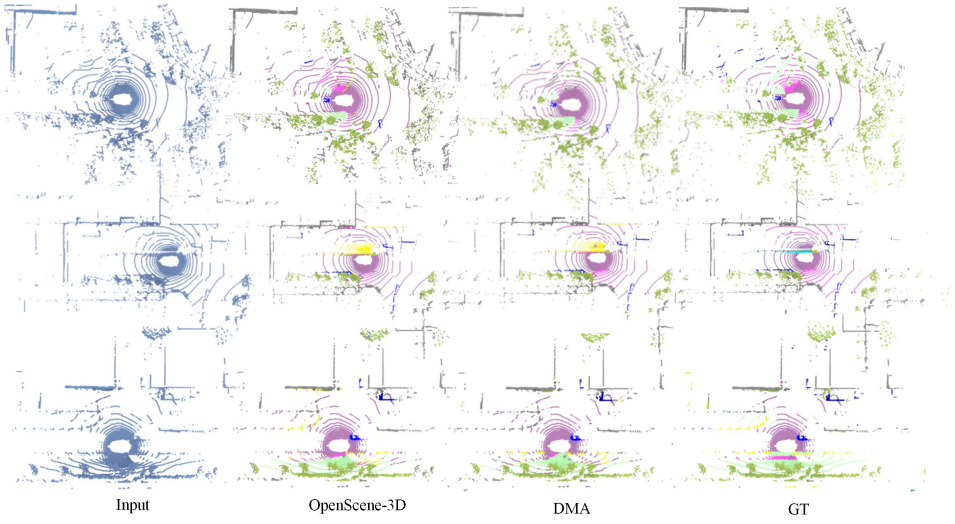

Fig. 5: Qualitative results of different methods on nuScenes dataset.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
2. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
3. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
4. Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al.: Recognize anything: A strong image tagging model. arXiv preprint arXiv:2306.03514 (2023)