A. Details of 3D CoT Dataset

Grounding Annotations. We collect large-scale 3D instance-text pairs (Figure 1) primarily from SceneVerse [8], MMScan [13], PQ3D [26], and Grounded 3D-LLM [3], as illustrated in Listing 3.

Navigation Annotations. We adapt instruction-following navigation data from R2R-CE [11], REVERIE-CE [15, 21], SRDF [20], and NavRAG [22] for the Habitat simulator environment [14, 17].

Grounding-Navigation Annotations. For open-vocabulary object grounding and navigation, we leverage instance-text annotations from HM3D [16, 23] and MP3D [2] that support the Habitat simulator [14, 17]. We further integrate synthetic data from HSSD [9], ProcTHOR-10K [5, 9], and ProcTHOR-Objaverse (AI2-THOR) [6, 10]. To generate trajectories for these instances:

- **Habitat Simulator:** For HM3D, MP3D, HSSD, and ProcTHOR-10K (Habitat), the simulator computes optimal trajectories based on the scene mesh, ensuring collision-free paths with minimal distance.
- **AI2-THOR Simulator:** For ProcTHOR-Objaverse (AI2-THOR), as the native simulator [10] lacks a sufficient ground-truth path computation function for our needs, we implemented an A* algorithm based on traversable BEV maps to generate shortest paths while avoiding obstacles.

Planning-Grounding-Navigation Annotations. These comprehensive samples are primarily sourced from SG3D [25] and Grounded 3D-LLM [3]. As shown in Listing 1, these annotations contain hierarchical instructions with multiple sub-goals. Each sub-goal specifies the target object category, its index in the instance point cloud (Figure 1), and the center coordinates for the instance.

Planning-Grounding Annotations. Scenes with planning annotations that are currently incompatible with the Habitat simulator, specifically ScanNet [4], 3RScan [18], and ARKitScenes [1], are utilized for planning and grounding tasks. While these samples lack navigational trajectories, they provide rich instance-level grounding and planning instructions.

Planning-Navigation Annotations. We incorporate data from VLN-Trans [24] (see Listing 2). In these samples, the sub-goal coordinates correspond to trajectory waypoints rather than specific object instances, focusing on path planning and navigation without explicit object grounding.

3D Scene Composition. The annotations described above span a wide range of environments, including posed RGB-D videos from real-world datasets (ScanNet [4], 3RScan [18], ARKitScenes [1]), high-quality real scans (HM3D [16], MP3D [2]), and synthetic scenes (HSSD [9], ProcTHOR-10K [5], ProcTHOR-Objaverse [7]).

Grounding Label Assignment. Unlike traditional 3D grounding tasks where all candidate objects are visible, the target object may be unobserved (i.e., lacking corresponding 3D tokens) during our online training. In such cases, the grounding output is assigned to a special `<grounding_none>` token. When tokens corresponding to the target instance exist within the patch or instance tokens, we optimize the model using a multi-label cross-entropy loss, treating the target tokens as positive samples and the rest as negative. Following g3D-LF [19] and Dynam3D [21], a 3D token is assigned to an instance based on its nearest neighbor in the ground-truth instance point cloud (available in HM3D, MP3D, ScanNet, 3RScan, and ARKitScenes). For scenes lacking instance point cloud annotations (e.g., HSSD, ProcTHOR-10K, and ProcTHOR-Objaverse), we only consider patch tokens within a 0.2m radius of the target instance center as positive samples.

B. Details of Real-world Mobile Manipulation

We validate the effectiveness of our D3D-VLP in real-world scenarios using the Hello Robot Stretch 3 mobile manipulator. The robot is equipped with a head-mounted Intel RealSense D435i RGB-D camera, which captures streaming posed RGB-D images. These streams are processed in real-time by the Dynam3D Encoder [21] to incrementally construct and update the Multi-level 3D Memory, ensuring the agent maintains a persistent and structured 3D scene representation during exploration.

For inference, we deploy the model on a remote workstation equipped with an NVIDIA RTX 4090 GPU and 64GB of RAM. The workstation handles the computationally intensive 3D-VLM reasoning and Dynamic 3D CoT generation, communicating with the robot via a local area network (LAN) over WiFi. Our deployment framework is adapted from DynaMem [12], which we extended to support our waypoint-based action space and local obstacle avoidance. The experiments are conducted in a physical home-like environment comprising a living room, kitchen, meeting room, and office. Crucially, to strictly evaluate zero-shot generalization, none of the objects or scene layouts in this environment are included in the training dataset.

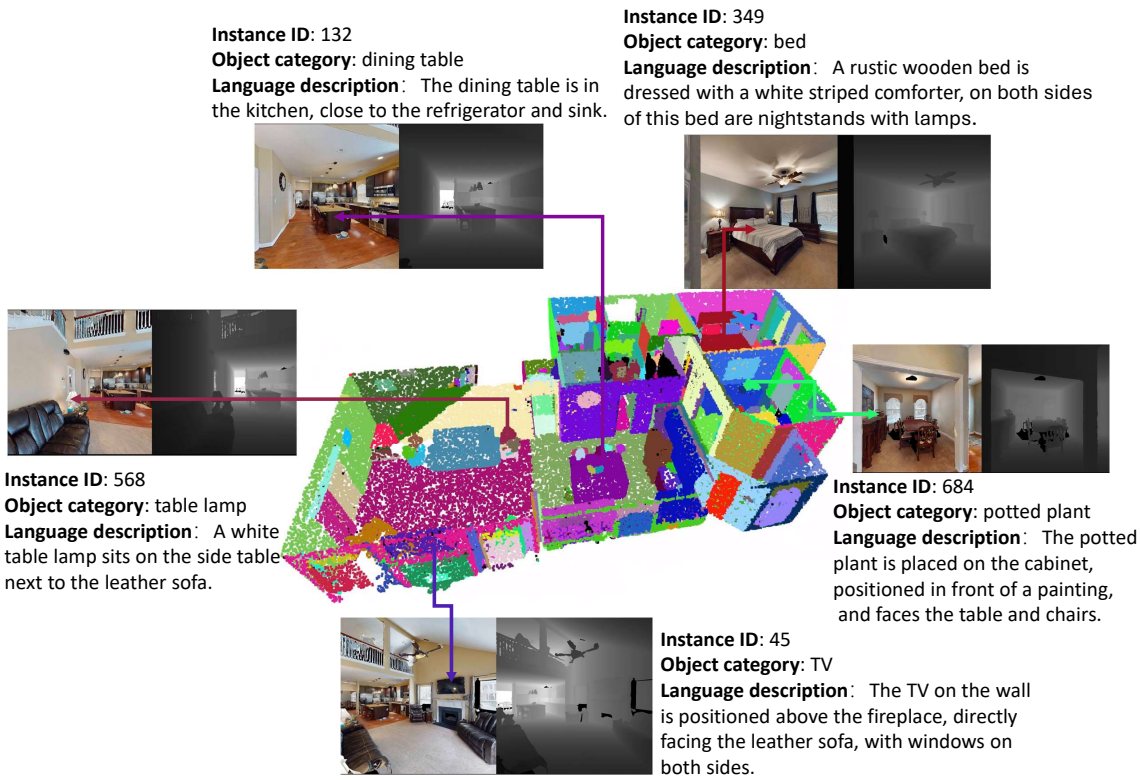

Figure 1. Demonstration of a 3D scene in the training data [8, 19]. Instance-level point clouds mark all instances with object categories and language descriptions.

Listing 1. **Planning-Grounding-Navigation Annotation.** An example showing the unified data structure and the corresponding prompt format.

```
// 1. JSON Annotation
{
  "scene_id": "hm3d/00378-DqJKU7YU7dA",
  "instruction": "Prepare for a shower.",
  "planning": [
    "1. Go to the shower containing a washcloth.",
    "2. Turn on the water to adjust the temperature.",
    "3. Take a towel from the rack stand to the left of the sink.",
    "4. Hang the towel on the shower curtain rod right by the shower."
  ],
  "instance_id": [[797], [797], [230], [515]],
  "instance_type": ["shower", "shower", "towel", "shower curtain rod"],
  // Coordinates truncated for display
  "instance_position": [
    [-6.554, -28.336, 0.863],
    [-6.554, -28.336, 0.863],
    [-2.407, -14.418, 1.189],
    [-3.649, -20.311, 0.981]
  ]
}

// 2. Input Prompt Trace
<|im_start|>system\n You are a helpful assistant<|im_end|\n
<|im_start|>user\n
<patch tokens>\n <instance tokens>\n <zone tokens>\n
The instruction: Prepare for a shower.\n
The history plans: 1. Go to the shower containing a washcloth.<shower_1>\n
The previous waypoints:<waypoint_4>,<waypoint_7>,<waypoint_9>,<waypoint_15>\n
The candidate waypoints: <waypoint_17><waypoint_18><waypoint_19>\n
Please give deep thinking plans. <|im_end|>\n<|im_start|>assistant\n
// 3. Output Trace
The next plans: 2. Turn on the water to adjust the temperature.\n
3. Take a towel from the rack stand to the left of the sink.\n
4. Hang the towel on the shower curtain rod right by the shower.\n
The grounded:target<shower_1>\n
The navigation action:waypoint<waypoint_19>reached the subgoal\n
<|im_end|>
```

Listing 2. **Planning-Navigation Annotation.** An example for vision-and-language navigation task with detailed step descriptions.


```
{
  "scene_id": "mp3d/PX4nDJXEHrG",
  "instruction": "Walk across patio into the house. Walk forward toward wall with stone tile. Walk past stone tile
     wall on left side. Walk past stair case. Stop at entrance to kitchen area.",
  "planning": [
    "1. Walk across patio into the house.",
    "2. Walk forward toward wall with stone tile.",
    "3. Walk past stone tile wall on left side.",
    "4. Walk past stair case.",
    "5. Stop at entrance to kitchen area."
  ],
  "habitat_start_position": [-11.546, 0.115, 4.632],
  "habitat_start_rotation": [0, 0.998, 0, -0.049],
  "instance_id": [[-10000],[-10000],[-10000],[-10000],[-10000]],
  "instance_type": [null,null,null,null,null],
  "instance_position": [
    [-11.559, -3.040, 0.115],
    [-11.066, -1.661, 0.115],
    [-12.035, 1.826, 0.115],
    [-13.690, 4.130, 0.115],
    [-13.619, 5.133, 0.115]
  ]
}

```


Listing 3. **Grounding-Navigation Annotation.** An example for open-vocabulary object grounding and navigation task without explicit intermediate planning steps.


```
{
  "scene_id": "hm3d/iigzG1rtanx",
  "instruction": "Please find the printer. The printer is next to the paper tray and is placed on desk.",
  "planning": "",
  "instance_id": [[96]],
  "instance_type": ["printer"],
  "instance_position": [
    [6.704, 1.147, 2.792]
  ]
}

```

References

[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. 1

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017. 1

[3] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, Runsen Xu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024. 1

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1

[5] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *NeurIPS*, 2022. Outstanding Paper Award. 1

[6] Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, et al. Spoc: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16238–16250, 2024. 1

[7] Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, et al. Spoc: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. In *CVPR*, pages 16238–16250, 2024. 1

[8] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2

[9] Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16384–16393, 2024. 1

[10] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 1

[11] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 104–120. Springer, 2020. 1

[12] Peiqi Liu, Zhanqiu Guo, Mohit Warke, Soumith Chintala, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation. In *CoRL 2024 Workshop on Mastering Robot Manipulation in a World of Abundant Data*. 1

[13] Ruiyuan Lyu, Tai Wang, Jingli Lin, Shuai Yang, Xiaohan Mao, Yilun Chen, Runsen Xu, Haifeng Huang, Chenming Zhu, Dahua Lin, and Jiangmiao Pang. Mmscan: A multi-modal 3d scene dataset with hierarchical grounded language annotations. In *arXiv*, 2024. 1

[14] Xavi Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Ruslan Partsey, Jimmy Yang, Ruta Desai, Alexander William Clegg, Michal Hlavac, Tiffany Min, Theo Gervet, Vladimír Vondruš, Vincent-Pierre Berges, John Turner, Oleksandr Maksymets, Zsolt Kira, Mrinal Kalakrishnan, Jitendra Malik, Devendra Singh Chaplot, Unnat Jain, Dhruv Batra, Akshara Rai, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars and robots, 2023. 1

[15] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020. 1

[16] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 1

[17] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *CVPR*, pages 9339–9347, 2019. 1

[18] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *ICCV*, 2019. 1

[19] Zihan Wang and Gim Hee Lee. g3d-lf: Generalizable 3d-language feature fields for embodied tasks. *arXiv preprint arXiv:2411.17030*, 2024. 1, 2

[20] Zun Wang, Jialu Li, Yicong Hong, Songze Li, Kunchang Li, Shoubin Yu, Yi Wang, Yu Qiao, Yali Wang, Mohit Bansal, et al. Bootstrapping language-guided navigation learning with self-refining data flywheel. In *The Thirteenth International Conference on Learning Representations*. 1

[21] Zihan Wang, Seungjun Lee, and Gim Hee Lee. Dynam3d: Dynamic layered 3d tokens empower vlm for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, 2025. 1

[22] Zihan Wang, Yaohui Zhu, Gim Hee Lee, and Yachun Fan. Navrag: Generating user demand instructions for embodied navigation through retrieval-augmented llm. *arXiv preprint arXiv:2502.11142*, 2025. 1

[23] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4927–4936, 2023. 1

[24] Yue Zhang and Parisa Kordjamshidi. Vln-trans: Translator for the vision and language navigation agent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13219–13233, 2023. 1

[25] Zhuofan Zhang, Ziyu Zhu, Junhao Li, Pengxiang Li, Tianxu Wang, Tengyu Liu, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Siyuan Huang, and Qing Li. Task-oriented sequential grounding and navigation in 3d scenes. *arXiv preprint arXiv:2408.04034*, 2024. 1

[26] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pages 188–206. Springer, 2024. 1