Navigation Instruction Generation with BEV Perception and Large Language Models
Supplementary Material

Sheng Fan, Rui Liu, Wenguan Wang†, and Yi Yang

ReLER, CCAI, Zhejiang University
https://github.com/FanScy/BEVInstructor

This document provides more details of BEVInstructor as follows:

- *Implementation Details of* BEVInstructor *(§A).* We introduce more implementation details of BEVInstructor and provide the pseudo-code for the training procedure.
- *Additional Quantitative and Qualitative Results (§B).* More quantitative and qualitative results are provided.
- *Discussion (§C).* We offer further discussion on the limitations, social impact, and future work of BEVInstructor.

A Implementation Details of BEVInstructor

BEV Encoding with 3D Detection. To establish a holistic 3D scene understanding, 3D detection is used to supervise the BEV encoding [6,15] in Eq. 4. For 3D detection, we adopt the 3D bounding boxes [8] of Matterport3D [2] following the same *seen*/*unseen* splits as R2R [1]. It contains 17 categories, covering common objects in daily life. The bipartite matching and the bounding box losses [6] are employed for detection. The BEV encoder is optimized by AdamW [10] for 500 epochs with a learning rate of 1×10^{-4}.

In Table S1, we provide the detailed results of the main categories. It is noted that we adhere to the same settings as those used in BEVFormer [6], *i.e.*, the shape of the BEV plane, the perception range, and the distribution of reference points (§3.5). For metrics, we use the 3D IoU-based mean average precision (mAP) with thresholds of 0.5. It demonstrates our depth consistency weight (Eq. 5) facilitates the BEV projection and obtains higher quality BEV representations for scene perception.

We also present the pseudo-code of the training procedure in Algorithm 1.

B Additional Quantitative and Qualitative Results

User Study. Due to the inherent limitations of evaluation metrics, the current metrics cannot fully reflect the performance of generated instructions [17]. To

† Corresponding author: *Wenguan Wang.*

Models	mAP↑	bed	table	door	sofa	chair	shelving	cabinet	plant	sink	cushion	monitor
BEVFormer [6]	23.37	33.75	32.71	15.27	29.55	30.15	7.59	26.07	21.96	21.63	17.23	21.14
BEVInstructor(Ours)	**25.42**	39.02	37.20	16.89	31.09	31.26	9.44	31.96	24.65	18.58	21.53	18.04

Table S1: Detailed results of the main categories on the *unseen* scenes.

Algorithm 1 The pseudo-code of training for BEVInstructor.

Arguments: Multi-view Image Features $\{\boldsymbol{F}_{t,k}\}_{k=1}^{K}$ with Orientation Angles $\{\delta_{t,k}\}_{k=1}^{K}$, Perspective Embedding $\boldsymbol{P}_t$, Action Embedding $\boldsymbol{a}_t$, BEV Embedding $\boldsymbol{B}_t$, Complete Observation Embedding $\boldsymbol{O}_t$, Instruction Tokens $\mathcal{X} = \{\boldsymbol{x}_l\}_{l=1}^{L}$, Landmark Tokens $\mathcal{X}^I$, the maximum iteration N, Perspective Encoder $\mathcal{E}^{\boldsymbol{P}}$, BEV Encoder $\mathcal{E}^{\boldsymbol{B}}$, Perspective-BEV Encoder $\mathcal{E}^{\boldsymbol{O}}$, MLLM $\mathcal{E}^{LLM}$.

Initialize $\mathcal{E}^{\boldsymbol{P}}$, $\mathcal{E}^{\boldsymbol{B}}$, $\mathcal{E}^{\boldsymbol{O}}$, $\mathcal{E}^{LLM}$
for iteration $i \in [1, \ldots, N]$ **do**
  Sample a pretraining task $\mathcal{T}$ from {*Landmarks*, *Instructions*}
  $[\boldsymbol{P}_t, \boldsymbol{a}_t] = \mathcal{E}^{\boldsymbol{P}}(\boldsymbol{F}_{t,k}, \delta_{t,k})$ ▷ Defined in Eq. 2, 3.
  $\boldsymbol{B}_t = \mathcal{E}^{\boldsymbol{B}}(\boldsymbol{F}_{t,a})$ ▷ Defined in Eq. 4, 5, 6.
  $\boldsymbol{O}_t = \mathcal{E}^{\boldsymbol{O}}(\boldsymbol{B}_t, [\boldsymbol{P}_t, \boldsymbol{a}_t])$ ▷ Defined in Eq. 7, 8.
  if $\mathcal{T}$ is *Landmarks* **then**
    $\mathcal{X}^I = \mathcal{E}^{LLM}(\boldsymbol{x}_{<l}, \boldsymbol{O}_{1:T})$ ▷ Defined in Eq. 10.
  else$\mathcal{T}$ is *Instructions*
    $\mathcal{X} = \mathcal{E}^{LLM}(\boldsymbol{x}_{<l}, \boldsymbol{O}_{1:T}, \mathcal{X}^I)$ ▷ Defined in Eq. 11.
  Update $\mathcal{E}^{\boldsymbol{P}}$, $\mathcal{E}^{\boldsymbol{B}}$, $\mathcal{E}^{\boldsymbol{O}}$, $\mathcal{E}^{LLM}$
return $\mathcal{E}^{\boldsymbol{P}}$, $\mathcal{E}^{\boldsymbol{B}}$, $\mathcal{E}^{\boldsymbol{O}}$, $\mathcal{E}^{LLM}$

more comprehensively reflect its performance, we conduct a set of human evaluation experiments. Specifically, 23 college students are invited to evaluate 100 instructions in total generated by various algorithms, including BEVInstructor, BT-speaker, EDrop-speaker, CCC-speaker, and Lana. They score each instruction based on its match with the navigation path using a scale from 0 to 5. The test paths for user study are sampled from REVERIE `val unseen`. As a result, BEVInstructor, with a score of 3.64, outperforms the other models, *i.e.*, BT-speaker 2.97, EDrop-speaker 3.11, CCC-speaker 3.24, and Lana 3.43.

More Examples. In Fig. S1, BEVInstructor successfully captures the crucial landmarks, *e.g.*, *fire extinguisher*, based on the perspective-BEV module, while other algorithms fail to provide more detailed information in their instructions.

C Discussion

Limitations. BEVInstructor outperforms existing methods across all datasets, *i.e.*, R2R [1], REVERIE [12], and UrbanWalk [4]. Despite notable progress, in comparison to human-annotated instructions, there exists considerable space for enhancing the diversity and accuracy of the instructions. In open-vocabulary settings, BEVInstructor continues to necessitate human intervention for the purpose of filtering and correction. Moreover, BEVInstructor currently does

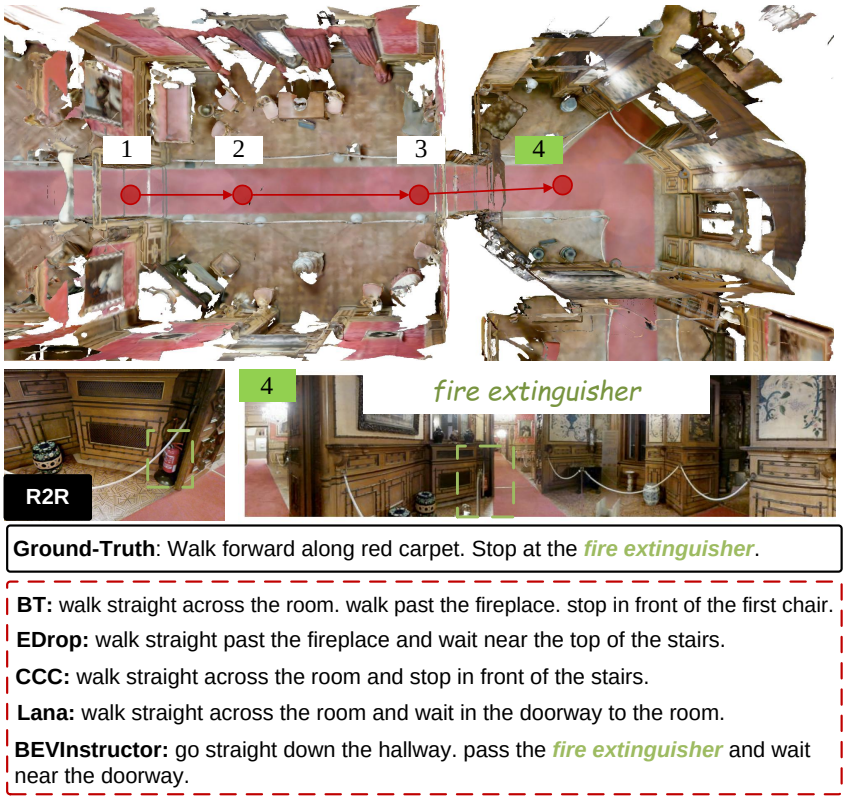

Fig. S1: Comparison results among Ground-Truth, BT-speaker [3], EDrop-speaker [13], CCC-speaker [14], Lana [16], and BEVInstructor for instruction generation on R2R [1] val unseen split. See §B for more details.

not incorporate safety factors, *e.g.*, warnings of dangerous areas, which are crucial for application in real-world scenarios.

Social Impact. The proposed framework for navigation instruction generation, incorporating MLLMs and BEV features, presents a pioneering contribution with substantial implications for social impact. Our approach not only achieves an impressive improvement in the performance, but also has stronger interpretability through outputting landmarks in the process of refinement. This approach can significantly enhance the trust between humans and agents during navigation, aligning more closely with human cognitive methods.

Future Work. BEVInstructor integrates perspective features and BEV features into a unified representation. Given that the BEV coordinate is consistent with the 3D coordinate, BEV naturally supports multi-sensor fusion [5, 7, 9, 11]. Future developments for BEVInstructor aim to expand this framework by incorporating the 3D world into the current MLLM via multi-sensor features, *e.g.*, LiDAR. This advancement will not only contribute to the robustness and

versatility of BEVInstructor but also elevate its efficacy in real-world scenarios. Furthermore, recognizing the importance of safety, future enhancements will focus on embedding navigational and environmental safety measures into the model.

References

1. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR (2018) 1, 2, 3
2. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from RGB-D data in indoor environments. In: 3DV (2017) 1
3. Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., Darrell, T.: Speaker-follower models for vision-and-language navigation. In: NeurIPS (2018) 3
4. Huang, Z., Shangguan, Z., Zhang, J., Bar, G., Boyd, M., Ohn-Bar, E.: Assister: Assistive navigation via conditional instruction generation. In: ECCV (2022) 2
5. Li, H., Sima, C., Dai, J., Wang, W., Lu, L., Wang, H., Zeng, J., Li, Z., Yang, J., Deng, H., et al.: Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 3
6. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV (2022) 1, 2
7. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., Tang, Z.: Bevfusion: A simple and robust lidar-camera fusion framework. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) NeurIPS (2022) 3
8. Liu, R., Wang, X., Wang, W., Yang, Y.: Bird's-eye-view scene graph for vision-language navigation. In: ICCV (2023) 1
9. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: ICRA (2023) 3
10. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 1
11. Ma, Y., Wang, T., Bai, X., Yang, H., Hou, Y., Wang, Y., Qiao, Y., Yang, R., Manocha, D., Zhu, X.: Vision-centric bev perception: A survey. arXiv preprint arXiv:2208.02797 (2022) 3
12. Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., van den Hengel, A.: Reverie: Remote embodied visual referring expression in real indoor environments. In: CVPR (2020) 2
13. Tan, H., Yu, L., Bansal, M.: Learning to navigate unseen environments: Back translation with environmental dropout. In: NAACL (2019) 3
14. Wang, H., Liang, W., Shen, J., Van Gool, L., Wang, W.: Counterfactual cycle-consistent learning for instruction following and generation in vision-language navigation. In: CVPR (2022) 3
15. Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., Liu, X., Lu, C., Lin, D., Pang, J.: Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In: CVPR (2024) 1

16. Wang, X., Wang, W., Shao, J., Yang, Y.: Lana: A language-capable navigator for instruction following and generation. In: CVPR (2023) 3
17. Zhao, M., Anderson, P., Jain, V., Wang, S., Ku, A., Baldridge, J., Ie, E.: On the evaluation of vision-and-language navigation instructions. In: EACL (2021) 1