PathPainter: Transferring the Generalization Ability of Image Generation Models to Embodied Navigation

Yijin Wang[1,2,*] **Yuru Tian**[1,2,*] **Xijie Huang**[1,2,*] **Weiqi Gai**[2,3]
Mo Zhu[1,2] **Xin Zhou**[2] **Yuze Wu**[1,2,†] **Fei Gao**[1,2,†]

Abstract: Bird's-eye-view (BEV) images have been widely demonstrated to provide valuable prior information for navigation. Given the global information provided by such views, two key challenges remain: how to fully exploit this information and how to reliably use it during execution. In this paper, we propose a navigation system that uses BEV images as global priors and is designed for ground and near-ground robotic platforms. The system employs an image generation model to interpret human intent from natural language, identify the target destination, and generate traversability masks. During execution, we introduce cross-view localization to align the robot's odometry with the BEV map and mitigate long-term drift in conventional odometry. We conduct extensive benchmark experiments to evaluate the proposed method and further validate it on a UAV platform. Using only a conventional local motion planner, the UAV successfully completes a 160-meter outdoor long-range navigation task. This work demonstrates how the world-understanding capabilities of foundation models can be transferred to embodied navigation, enabling robots to benefit from the strong generalization ability of existing image generation models.

Keywords: Robot Navigation, BEV Prior, Foundation Models, Cross-view Localization, Image Generation Model

1 Introduction

BEV images from aerial or satellite perspectives provide robots with valuable global priors for outdoor navigation, including road topology, open areas, obstacle layouts, and spatial relationships between targets and surrounding structures [1, 2, 3]. These priors are particularly important for natural-language-guided navigation, where robots must ground language instructions in complex scenes and infer feasible routes [4, 5]. As a result, bird's-eye-view information has been widely used in air-ground collaborative navigation [6, 7, 8, 9], satellite-map-based navigation [1, 2, 10], and high-low altitude collaborative navigation [11, 12, 13].

Existing BEV-based navigation systems still struggle to convert aerial observations into executable navigation priors. Many methods compress BEV images into semantic graphs, topological maps, or category-level representations for efficient planning [6, 7]. Although compact, these representations may discard geometric details and visual cues needed for navigation, such as road boundaries, passage width, obstacle layout, and open-space continuity.

Moreover, traversability, i.e., whether a region can support feasible robot movement, is often approximated by human-predefined semantic classes, especially roads [6]. Such road-centric priors are effective in structured scenes, but fail to capture diverse traversable regions in open outdoor environments, where feasible navigation may involve sidewalks, trails, off-road routes, and movement across plazas, beaches, and other open spaces.

[1]Zhejiang University. [2]Differential Robotics. [3]Beihang University. *Equal contribution. †Corresponding authors: wuyuze000@zju.edu.cn, fgaoaa@zju.edu.cn.

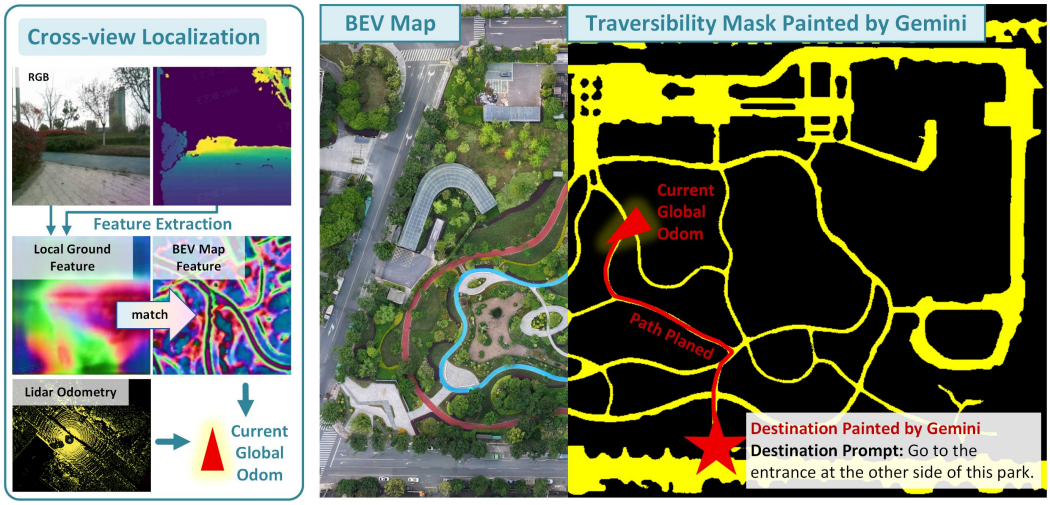

Figure 1: Overview of the Navigation System. **Left:** Cross-view localization extracts embeddings from local ground features reconstructed from RGB-D observations and matches them with feature embeddings from the BEV map to estimate the robot's global odometry. **Right:** Given the destination prompt and the BEV map, the image generation model marks the target region with a generated star marker and produces a traversability mask. The red path is obtained by running A^* on the mask.

Recent foundation models have demonstrated strong generalization capabilities in visual, language, and spatial understanding. In particular, image generation models [14, 15] encode rich priors about spatial layout, object relations, and navigability cues. Prior work has shown that visual understanding tasks can be reformulated as image generation problems [16].

These observations motivate our central question: can BEV navigation be reformulated as an image generation problem, allowing image generation models to produce dense navigation priors for target grounding, traversability estimation, and global path planning, and further convert them into paths that are structured, searchable, and executable by real robots?

In this paper, we propose PathPainter, a path-planning framework for natural-language-conditioned BEV navigation, as shown in Fig. 1. Given a BEV image, the robot's start position, and a language instruction, PathPainter infers the target region, generates a navigation-oriented traversability mask, and applies A^* search to obtain an executable global path. We further integrate PathPainter into a real-world long-range navigation system that combines cross-view localization, BEV-based global planning, and local motion execution. Cross-view localization aligns the robot trajectory with the BEV map without relying on precise GPS/RTK localization.

PathPainter demonstrates a practical way to transfer the generalization ability of image generation models to embodied navigation through BEV prior interpretation. Our main contributions are: (1) We propose PathPainter, which transfers generative vision priors to natural-language-conditioned BEV navigation through destination inference, traversability-mask generation, and search-based planning. (2) We develop a real-world long-range navigation system combining cross-view localization, BEV global planning, and local motion planning for near-ground outdoor robots. (3) We validate PathPainter on real-world navigation and path-generation benchmarks, demonstrating its robustness. Code will be released at `https://github.com/E-hash-42/PathPainter`.

2 Related Work

2.1 Navigation with Priors

Compared with onboard local observations, prior information provides larger-scale environmental structure and potential goal locations, reducing exploration cost in long-range navigation. Existing

works use priors such as OpenStreetMap [4, 17], abstract topological maps [18], and aerial maps [19, 20] for global planning, subgoal selection, or local planning constraints. To use these priors for navigation, prior works often abstract them into road graphs, topological nodes, or semantic regions, and study how to align onboard observations with the global prior during execution. While useful for planning and localization, such abstractions may discard fine-grained geometry, visual appearance, and local traversability cues in aerial images. In contrast, our work directly extracts traversable regions from aerial images for A*-based planning on the resulting masks, and introduces cross-view localization to align robot odometry with the aerial prior during execution.

2.2 Navigation with Foundation Models

Recent works have introduced foundation models into navigation, enabling robots to understand human instructions and perform goal, object, and language-guided navigation in complex environments. These methods mainly follow two paradigms: vision-language models directly predict actions or high-level waypoints from instructions [21, 22], while video generation models synthesize instruction-conditioned future videos and infer executable actions from them [23, 24]. Some methods further combine both paradigms [25]. Overall, these approaches demonstrate the semantic understanding and cross-scene generalization of foundation models, advancing open-world navigation.

3 Method

3.1 Pipeline

As shown in Fig. 2, this paper proposes a bird's-eye-view-based navigation planning framework. The system takes an aerial orthophoto and a natural-language instruction as input, outputs an executable navigation path, and deploys it on a real robotic platform for autonomous execution.

The overall pipeline consists of two levels: high-level path generation and low-level motion execution. At the high level, the system first uses an image generation model to semantically understand the BEV map, draw the traversability mask, and extract a global path based on the start and goal positions. At the low level, the system maps the global path to the robot's local coordinate frame through cross-view localization, and combines LiDAR odometry with a robust local planner [26] for real-time obstacle avoidance and trajectory tracking.

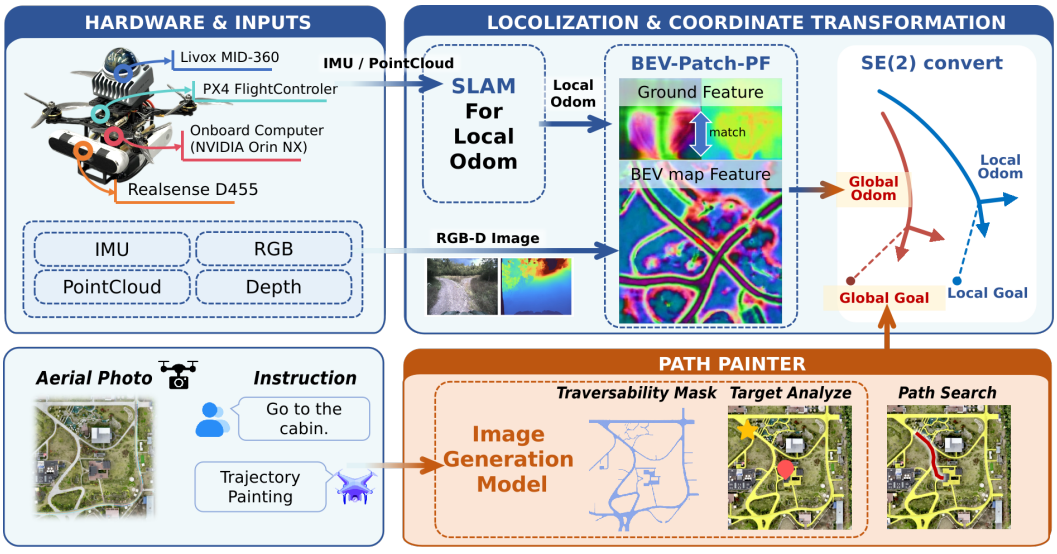

Figure 2: Pipeline of our navigation system.

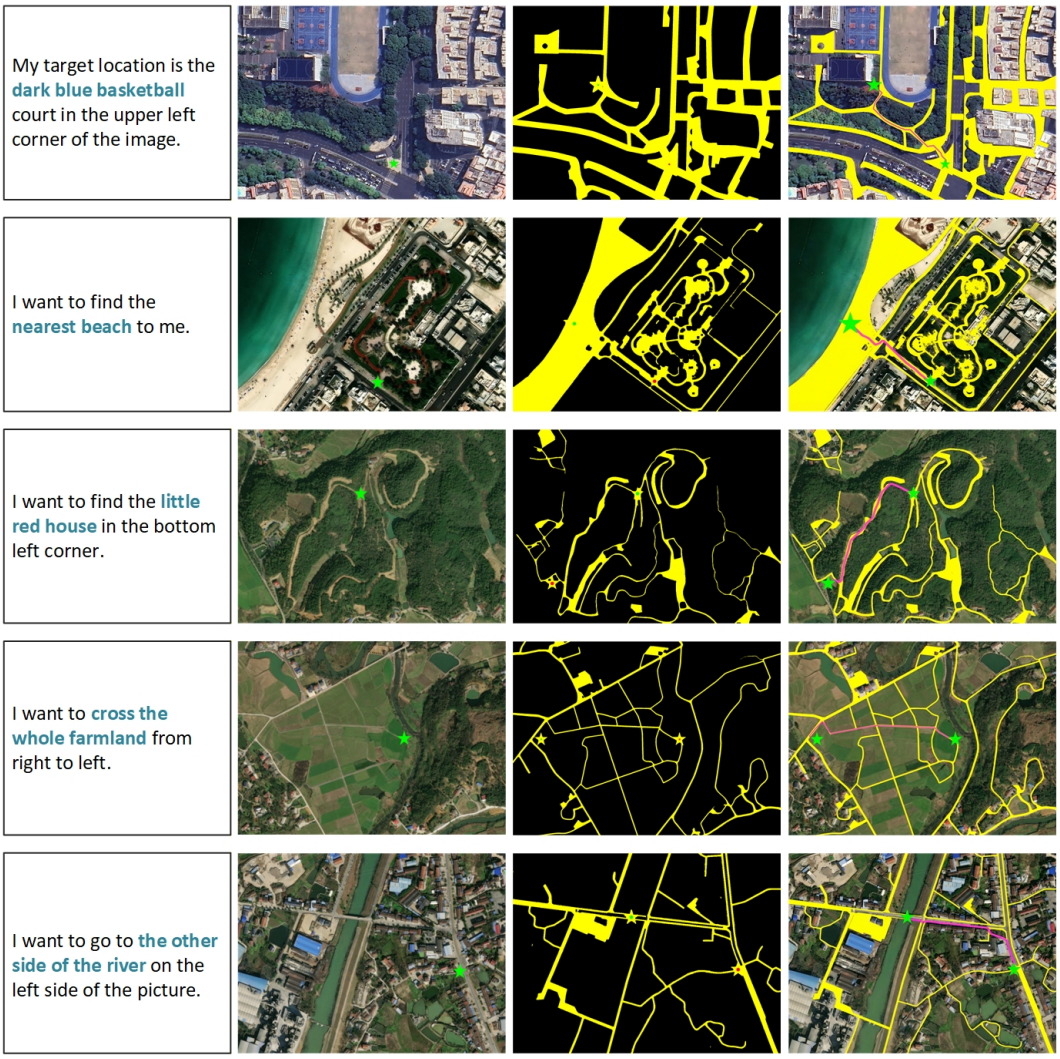

Figure 3: Workflow of PathPainter. **Column 1:** natural-language destination query. **Column 2:** original map with the current robot position. **Column 3:** Traversability mask. **Column 4:** final planning result, where the generated traversability mask, predicted goal position, and planned path are overlaid on the original map.

This hierarchical design decouples high-level semantic reasoning from low-level motion control, enabling stable navigation in complex and unstructured outdoor environments.

3.2 Extracting Executable Paths via Image-to-Image Generation

We formulate the path planning problem as an image-to-image generation process, thereby leveraging the spatial understanding and semantic reasoning capabilities of image generation models to transform high-level map understanding into an executable navigation path. As shown in Fig. 3, the image generation model is responsible for two main tasks: goal position prediction and traversability-mask segmentation. Specifically, we first mark the robot's current position on the BEV map with a green star. Conditioned on this starting point and the natural-language instruction, the image generation model is prompted to infer and draw the target position.

Meanwhile, the system employs carefully designed prompts to guide the image generation model to generate an image mask representing the traversable area, referred to as the traversability mask. This mask explicitly encodes walkable regions in the scene, such as roads, sidewalks, and open

spaces, while suppressing non-traversable areas such as buildings, vegetation, and obstacles. For road regions that are partially occluded by trees but remain semantically continuous, the model can preserve their connectivity based on global contextual reasoning. This process can therefore be regarded as a mapping from raw visual input to a navigable-space representation.

Finally, we further extract a feasible path connecting the start and goal positions from the generated traversability mask. Specifically, A* search is performed on the binary mask corresponding to the traversable region. To encourage the planned path to lie near the center of roads, we introduce a boundary-distance-based penalty term into the A* cost function: path points farther from the boundary of the traversable region are assigned lower costs. This encourages the resulting path to stay away from boundaries, thereby improving both safety and executability.

3.3 Cross-view Localization

To execute the BEV global path, we use BEV-Patch-PF [27] to align the robot's local odometry frame with the BEV map frame. It estimates the robot pose on the BEV map as $\mathbf{x}_t^M = (x_t^M, y_t^M, \theta_t^M)$, where M denotes the map frame.

Given the global path $\mathcal{P}^M = \{\mathbf{p}_1^M, \ldots, \mathbf{p}_K^M\}$, we select a look-ahead waypoint $\mathbf{p}_k^M$ according to the current global odometry and transform it into the local odometry frame O:

$$\tilde{\mathbf{p}}_k^O = \mathbf{T}_M^O \tilde{\mathbf{p}}_k^M, \tag{1}$$

where $\mathbf{T}_M^O$ is computed from the cross-view localization result and the current odometry estimate, and $\tilde{\mathbf{p}}$ denotes the homogeneous waypoint coordinate. The transformed waypoint is sent to the local planner as a short-horizon navigation goal, allowing high-rate local planning while periodically correcting long-term odometry drift.

4 Experiment

4.1 Evaluating Path Generation Methods

It is important to note that our goal is not pure semantic segmentation or complete road-topology reconstruction. Instead, we target **traversability-prior generation for navigation**. This task requires (1) recognizing visible roads, (2) reasoning about continuity under occlusion and plausible traversability in ambiguous regions, and (3) capturing diverse forms of traversability beyond road-like structures. Although models such as SAM 3.1 [28] and road-topology models such as RNGDet++ [29] and SAMRoad [30] are strong within their own design objectives, their outputs are not necessarily optimal for downstream path planning. Nevertheless, we benchmark these representative methods from two perspectives: their ability to produce stable visible-road priors and their utility for downstream path planning.

Visible-road segmentation benchmark. We evaluate zero-shot visible-road segmentation on DeepGlobe [31], UAVid [32], and VDD [33]. All methods are tested without task-specific training, fine-tuning, or manual correction. For image generation models, we follow the original prompt format [16] with a target keyword `road`, and convert outputs into binary masks. For open-vocabulary segmentation baselines, including SAM 3.1 and Text2Seg [34], we use `road` as the class prompt.

Table 1 shows that Gemini achieves strong recall across all three datasets, indicating that it tends to produce more continuous road regions. This property is beneficial for subsequent skeletonization and A* search. SAM 3.1 achieves higher precision and much lower inference time, showing that it is an efficient and reliable visible-road segmenter. However, it is more conservative and does not naturally perform occlusion completion or task-level traversability reasoning.

Downstream path-planning benchmark. To further evaluate whether these priors are useful for navigation, we construct a start-goal path-planning benchmark on CityScale [35] and GlobalScale [36]. Both datasets contain high-resolution 2048×2048 aerial images with ground-truth road annotations. For each image, we randomly sample 1000 start-goal pairs and evaluate whether each

Dataset	Method	IoU	Prec.	Rec.	F1	Time (s)
DeepGlobe [31]	GPT [15]	0.288	0.331	0.672	0.439	48.19
	Gemini [14]	**0.490**	0.613	**0.738**	**0.657**	47.68
	SAM 3.1 [28]	0.375	**0.659**	0.512	0.602	**0.17**
	Text2Seg [34]	0.063	0.100	0.527	0.148	0.46
UAVid [32]	GPT [15]	0.403	0.541	0.648	0.583	91.41
	Gemini [14]	**0.629**	0.713	**0.838**	0.754	41.12
	SAM 3.1 [28]	0.604	**0.848**	0.691	**0.781**	**0.17**
	Text2Seg [34]	0.373	0.514	0.500	0.539	0.48
VDD [33]	GPT [15]	0.248	0.281	0.625	0.454	154.40
	Gemini [14]	0.530	0.576	**0.843**	0.736	50.65
	SAM 3.1 [28]	**0.580**	**0.750**	0.732	**0.757**	**0.18**
	Text2Seg [34]	0.268	0.380	0.445	0.555	0.45

Table 1: **Visible-road segmentation benchmark.** All methods segment only the road category.

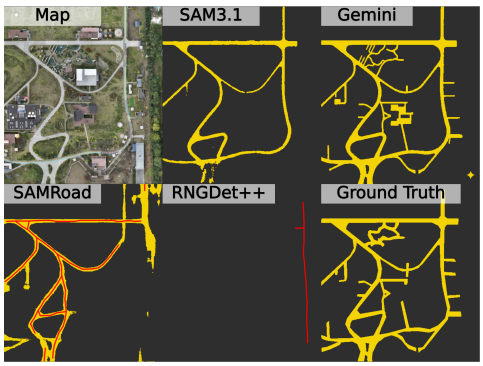

Figure 4: Real-world test on highly out-of-distribution scenes.

Method	In-domain (CityScale) [35]			OOD (Global-Scale) [36]			Time (s)
	Succ.	Valid.	Len.	Succ.	Valid.	Len.	
Gemini [14]	0.902	0.932	1.007	**0.766**	0.904	1.071	61.60
Gemini-Direct [14]	0.280	0.910	1.242	0.293	0.828	**1.038**	86.58
SAMRoad [30]	0.853	**0.994**	1.095	0.339	**0.975**	1.164	**9.34**
RNGDet++ [29]	**0.972**	0.969	**1.005**	0.252	0.949	1.148	39.95
SAM 3.1 [28]	0.018	0.016	0.829	0.042	0.040	0.887	0.266

Table 2: **Downstream path-planning benchmark.** All methods are evaluated only on the road category. Valid. (Validity) and Len. (Length Ratio) are computed only over successful samples. The Len. metric indicates better performance when its value is closer to 1.

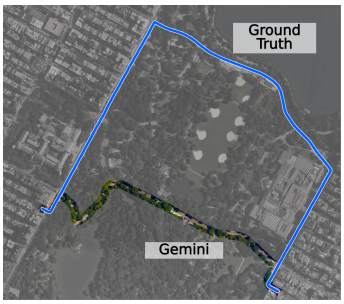

Figure 5: Gemini predicts roads unlabeled in the ground truth.

method can support path planning between them. CityScale is used as the in-domain road-only benchmark, while the out-of-domain (OOD) split of Global-Scale is used to evaluate cross-domain generalization. For fair comparison, all methods are evaluated only on the road category: their outputs are converted into binary road traversability masks and passed to the same A* planner. Although our full system targets broader traversability, this benchmark is restricted to roads because existing public datasets provide road-only annotations. For Gemini, the prompt also targets road traversability, while encouraging the prior to preserve global topological connectivity.

We report *Succ.* as the fraction of sampled start-goal pairs where the planner finds a connected path on the predicted traversability mask. For successful cases, *Valid.* measures the path fraction within the ground-truth traversable region, and *Len.* denotes the predicted-to-ground-truth shortest-path length ratio, where values closer to 1 are preferred. Average inference time is also reported.

As shown in Table 2, RNGDet++ and SAMRoad perform well on the in-domain CityScale split, but their success rates drop markedly on the Global-Scale OOD split due to missing or disconnected road segments under domain shift. SAM 3.1 further suffers from fragmented road segmentation and fails to recognize narrow paths in satellite images. As illustrated in Fig. 4, these benchmark methods are often impractical in real-world scenes. In contrast, Gemini achieves higher OOD reachability and a more stable length ratio, showing stronger robustness for global route generation. Its lower validity is partly caused by annotation mismatch: as shown in Fig. 5, Gemini may use visually plausible road-like regions that preserve connectivity but are unlabeled as roads, leading to lower validity. Its slight OOD success-rate drop mainly comes from forest coverage, tree occlusions, and cluttered road appearances. Nevertheless, Gemini provides more reliable global path candidates than graph-based baselines with only a modest validity trade-off.

We also test Gemini-Direct, where the model directly generates a path-oriented mask between the start and goal, rather than a complete traversability prior. We then apply the same A* planner to this

path-oriented mask for fair comparison. Its lower success rate and unstable path quality show that direct path-mask generation is less reliable, often producing invalid shortcuts across non-traversable regions. Thus, our system uses the prior-generation-and-search pipeline.

4.2 Real-world Results

As illustrated in Fig. 2, we evaluate our system through real-world deployment on a quadrotor UAV platform. The platform integrates an Intel RealSense D455 camera for cross-view localization and a Livox Mid-360 LiDAR for state estimation and local mapping. To ensure near-ground flight safety and approximate the viewpoint of ground mobile platforms, the UAV is constrained to a fixed altitude of 1 m. A prior BEV map is constructed by capturing aerial imagery with a DJI Neo drone and processing it via WebODM to generate a georeferenced orthomosaic (TIFF), which contains scale information and serves as the BEV map. The entire pipeline runs onboard an NVIDIA Orin NX (16 GB) edge computing platform. For cross-view localization, we directly use the feature extractor from BEV-Patch-PF [27] without fine-tuning on any of the experimental scenes. Due to onboard compute constraints, the cross-view localization module outputs global odometry at 1 Hz. We decouple this from the local planning stack: the 1 Hz global odometry is used exclusively for periodic drift correction and local goal updates, while a local motion planner operates independently at 10 Hz, using 200 Hz odometry obtained by fusing FAST-LIO2 [37] with IMU measurements for real-time obstacle avoidance and trajectory tracking.

We evaluate the complete system in 7 real-world scenarios with 21 navigation routes under diverse layouts and weather conditions. Across all trials, destination grounding and traversable-area generation are performed online using Gemini. This section reports three representative experiments from two scenarios, as shown in Figs. 6 and 7. Each trial evaluates the full pipeline from natural-language instruction to physical execution. The system succeeds in 15 of 21 routes, achieving a 71.4% end-to-end success rate. Two trials fail due to disconnected traversability masks, and four fail due to cross-view localization errors.

Figure 6: **Experiment 1.** Navigation in a park. The initial global pose estimate contains relatively large errors, making FAST-LIO2 alone insufficient for long-range navigation.

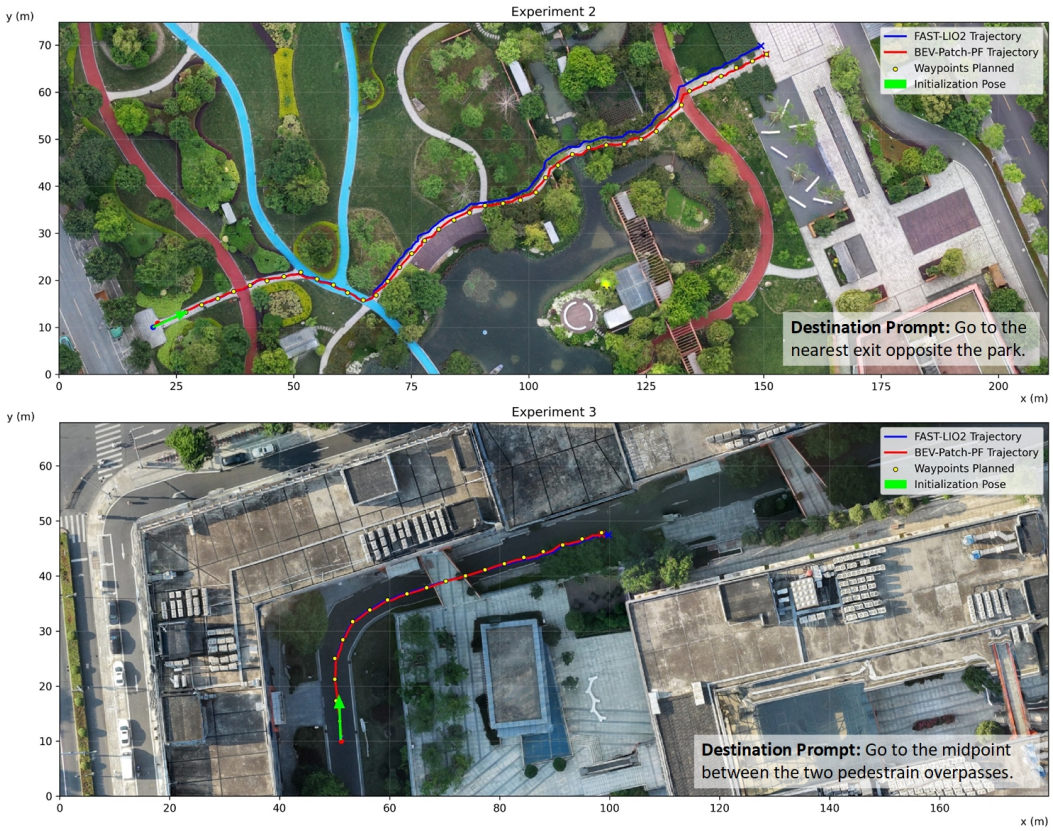

Figure 7: **Experiment 2:** Navigation in a park. Even with an accurate initial pose, unacceptable drift occurs during long-range navigation. **Experiment 3:** Navigation in structurally complex buildings.

Experiment 1 (Fig. 6) is conducted in an urban park without annotated pathways in commercial mapping services, highlighting the need for online environmental understanding and path planning. Raw FAST-LIO2 exhibits large initial pose errors and long-term drift, making it unsuitable for direct long-range navigation. In contrast, the cross-view module provides continuous global localization and dynamic correction of the local navigation goal. Without reliable RTK/GPS ground truth, we qualitatively validate the trajectory using synchronized RGB and aerial footage with manually annotated waypoints, showing that the UAV follows the planned corridor.

Experiments 2 and 3 (Fig. 7) demonstrate language-conditioned navigation capabilities. Experiment 3 is conducted in an industrial campus, where large GPS biases and fluctuations make reliable global localization challenging. The system interprets natural-language commands to extract semantic destination cues, generates feasible paths, and executes missions.

5 Conclusion and Limitations

We propose a BEV-based navigation planning framework driven by an image generation model. By formulating both destination selection and traversability-mask generation as image generation tasks, and leveraging cross-view localization for reliable global odometry, our approach transfers the strong generalization capability of image generation models to long-range outdoor navigation.

Several limitations remain. The system depends on accurate aerial priors, as misaligned or non-orthorectified maps can degrade performance. Moreover, 2D orthomosaics lack elevation information, causing ambiguity in uneven or multi-level scenes. Finally, onboard compute limits restrict cross-view localization to 1 Hz, which may limit performance during high-speed motion.

References

[1] M. Elnoor, K. Weerakoon, G. Seneviratne, R. Xian, T. Guan, M. K. M. Jaffar, V. Rajagopal, and D. Manocha. Robot navigation using physically grounded vision-language models in outdoor environments. *arXiv preprint arXiv:2409.20445*, 2024.

[2] C. Klammer and M. Kaess. Bevloc: Cross-view localization and matching via birds-eye-view synthesis. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5656–5663. IEEE, 2024.

[3] J. Zhang, H. Dong, J. Yang, J. Liu, S. Huang, K. Li, X. Tang, X. Wei, and X. You. Dual-bev nav: Dual-layer bev-based heuristic path planning for robotic navigation in unstructured outdoor environments. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8872–8879. IEEE, 2025.

[4] J. Lee, T. Miyanishi, S. Kurita, K. Sakamoto, D. Azuma, Y. Matsuo, and N. Inoue. City-nav: A large-scale dataset for real-world aerial navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5912–5922, 2025.

[5] C. Huang, O. Mees, A. Zeng, and W. Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023.

[6] F. Cladera, Z. Ravichandran, J. Hughes, V. Murali, C. Nieto-Granda, M. A. Hsieh, G. J. Pappas, C. J. Taylor, and V. Kumar. Air-ground collaboration for language-specified missions in unknown environments. *IEEE Transactions on Field Robotics*, 2025.

[7] H. Liu, Z. Ma, Y. Li, J. Sugihara, Y. Chen, J. Li, and M. Zhao. Hierarchical language models for semantic navigation and manipulation in an aerial-ground robotic system. *Advanced Intelligent Systems*, 8(2):e202500640, 2026. doi:https://doi.org/10.1002/aisy.202500640. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/aisy.202500640.

[8] Z. Li, R. Mao, N. Chen, C. Xu, F. Gao, and Y. Cao. Colag: A collaborative air-ground framework for perception-limited ugvs' navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16781–16787. IEEE, 2024.

[9] J. Deng, J. Liu, and J. Hu. Tightly-coupled air-ground collaborative system for autonomous ugv navigation in gps-denied environments. *Drones*, 9(9):614, 2025.

[10] Y. Huang, H. Dugmag, T. D. Barfoot, and F. Shkurti. Stochastic planning for asv navigation using satellite images. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 1055–1061. IEEE, 2023.

[11] R. Wu, Y. Zhang, J. Chen, L. Huang, S. Zhang, X. Zhou, L. Wang, and S. Liu. Aeroduo: Aerial duo for uav-based vision and language navigation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2576–2585, 2025.

[12] I. Munasinghe, A. Perera, and R. C. Deo. A comprehensive review of uav-ugv collaboration: Advancements and challenges. *Journal of Sensor and Actuator Networks*, 13(6):81, 2024.

[13] Y. Zhang, H. Yan, D. Zhu, J. Wang, C.-H. Zhang, W. Ding, X. Luo, C. Hua, and M. Q.-H. Meng. Air-ground collaborative robots for fire and rescue missions: Towards mapping and navigation perspective. *arXiv preprint arXiv:2412.20699*, 2024.

[14] Google. Introducing nano banana pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, Nov. 2025. Google Blog. Accessed: 2026-04-27.

[15] OpenAI. Chatgpt images 2.0 is now available. https://openai.com/zh-Hans-CN/index/introducing-chatgpt-images-2-0/, Apr. 2026. Accessed: 2026-04-27.

[16] V. Gabeur, S. Long, S. Peng, P. Voigtlaender, S. Sun, Y. Bao, K. Truong, Z. Wang, W. Zhou, J. T. Barron, K. Genova, N. Kannen, S. Ben, Y. Li, M. Guo, S. Yogin, Y. Gu, H. Chen, O. Wang, S. Xie, H. Zhou, K. He, T. Funkhouser, J.-B. Alayrac, and R. Soricut. Image generators are generalist vision learners, 2026. URL https://arxiv.org/abs/2604.20329.

[17] A. Li, Z. Wang, J. Zhang, M. Li, Y. Qi, Z. Chen, Z. Zhang, and H. Wang. Urbanvla: A vision-language-action model for urban micromobility. *arXiv preprint arXiv:2510.23576*, 2025.

[18] A. H. Tan, A. Fung, H. Wang, and G. Nejat. Mobile robot navigation using hand-drawn maps: A vision language model approach. *IEEE Robotics and Automation Letters*, 2025.

[19] C. Moore, S. Mitra, N. Pillai, M. Moore, S. Mittal, C. Bethel, and J. Chen. Ura*: Uncertainty-aware path planning using image-based aerial-to-ground traversability estimation for off-road environments. *arXiv preprint arXiv:2309.08814*, 2023.

[20] S. Shair, J. Chandler, V. Gonzalez-Villela, R. M. Parkin, and M. Jackson. The use of aerial images and gps for mobile robot waypoint navigation. *IEEE/ASME Transactions On Mechatronics*, 13(6):692–699, 2008.

[21] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and H. Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024.

[22] Y. Wu, M. Zhu, X. Li, Y. Du, Y. Fan, W. Li, Z. Han, X. Zhou, and F. Gao. Vla-an: An efficient and onboard vision-language-action framework for aerial navigation in complex environments. *arXiv preprint arXiv:2512.15258*, 2025.

[23] H. Zhang, S. Liang, L. Chen, Y. Li, Y. Xu, Y. Zhong, F. Zhang, and H. Li. Sparse video generation propels real-world beyond-the-view vision-language navigation. *arXiv preprint arXiv:2602.05827*, 2026.

[24] X. Huang, W. Gai, T. Wu, C. Wang, Z. Liu, X. Zhou, Y. Wu, and F. Gao. Navdreamer: Video models as zero-shot 3d navigators. *arXiv preprint arXiv:2602.09765*, 2026.

[25] J. Hu, J. Chen, H. Bai, M. Luo, S. Xie, Z. Chen, F. Liu, Z. Chu, X. Xue, B. Ren, et al. Astranav-world: World model for foresight control and consistency. *arXiv preprint arXiv:2512.21714*, 2025.

[26] X. Zhou, Z. Wang, H. Ye, C. Xu, and F. Gao. Ego-planner: An esdf-free gradient-based local planner for quadrotors. *IEEE Robotics and Automation Letters*, 6(2):478–485, 2020.

[27] D. Lee, J. Quattrociocchi, C. Ellis, R. Rana, A. Adkins, A. Uccello, G. Warnell, and J. Biswas. Bev-patch-pf: Particle filtering with bev-aerial feature matching for off-road geo-localization. *arXiv preprint arXiv:2512.15111*, 2025.

[28] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollár, N. Ravi, K. Saenko, P. Zhang, and C. Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

[29] Z. Xu, Y. Liu, Y. Sun, M. Liu, and L. Wang. Rngdet++: Road network graph detection by transformer with instance segmentation and multi-scale features enhancement. *IEEE Robotics and Automation Letters*, 8(5):2991–2998, 2023.

[30] C. Hetang, H. Xue, C. Le, T. Yue, W. Wang, and Y. He. Segment anything model for road network graph extraction. arxiv. *arXiv preprint arXiv:2403.16051*, 2024.

[31] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 172–181, 2018.

[32] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS journal of photogrammetry and remote sensing*, 165:108–119, 2020.

[33] A. Yeshchenko, J. Mendling, C. Di Ciccio, and A. Polyvyanny. Vdd: A visual drift detection system for process mining. 2020.

[34] J. Zhang, Z. Zhou, G. Mai, M. Hu, Z. Guan, S. Li, and L. Mu. Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models. *arXiv preprint arXiv:2304.10597*, 2023.

[35] S. He, F. Bastani, S. Jagwani, M. Alizadeh, H. Balakrishnan, S. Chawla, M. M. Elshrif, S. Madden, and M. A. Sadeghi. Sat2graph: Road graph extraction through graph-tensor encoding. In *European Conference on Computer Vision*, pages 51–67. Springer, 2020.

[36] P. Yin, K. Li, X. Cao, J. Yao, L. Liu, X. Bai, F. Zhou, and D. Meng. Towards satellite image road graph extraction: A global-scale dataset and a novel method. *arXiv preprint arXiv:2411.16733*, 2024.

[37] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang. Fast-lio2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics*, 38(4):2053–2073, 2022.