Caption-Guided Retrieval System for Cross-Modal Drone Navigation

Lingfeng Zhang*
Tsinghua University
Xiaomi EV
zlf25@mails.tsinghua.edu.cn

Erjia Xiao*
HKUSTGZ
exiao469@connect.hkust-gz.edu.cn

Yuchen Zhang
Georgia Institute of Technology
Xiaomi EV
yzhang4224@gatech.edu

Haoxiang Fu
National University of Singapore
e1127454@u.nus.edu

Ruibin Hu
The Chinese University of Hong Kong
1155223884@link.cuhk.edu.hk

Yanbiao Ma
Renmin University of China
ybma1998xidian@gmail.com

Wenbo Ding
Tsinghua University
ding.wenbo@sz.tingshua.edu.cn

Long Chen, Hangjun Ye, Xiaoshuai Hao†
Xiaomi EV
{chenlong37,yehangjun,haoxiaoshuai}@xiaomi.com

Abstract

Cross-modal drone navigation remains a challenging task in robotics, requiring efficient retrieval of relevant images from large-scale databases based on natural language descriptions. The RoboSense 2025 Track 4 challenge addresses this challenge, focusing on robust, natural language-guided cross-view image retrieval across multiple platforms (drones, satellites, and ground cameras). Current baseline methods, while effective for initial retrieval, often struggle to achieve fine-grained semantic matching between text queries and visual content, especially in complex aerial scenes. To address this challenge, we propose a two-stage retrieval refinement method: ***Caption-Guided Retrieval System (CGRS)*** *that enhances the baseline coarse ranking through intelligent reranking. Our method first leverages a baseline model to obtain an initial coarse ranking of the top 20 most relevant images for each query. We then use Vision-Language-Model (VLM) to generate detailed captions for these candidate images, capturing rich semantic descriptions of their visual content. These generated captions are then used in a multimodal similarity computation framework to perform fine-grained reranking of the original text query, effectively building a semantic bridge between the visual content and natural language descriptions. Our approach significantly improves upon the baseline, achieving a consistent 5% improvement across all key metrics (Recall@1, Recall@5, and Recall@10). Our approach win* ***TOP-3*** *in the challenge, demonstrating the practical value of our semantic refinement strategy in real-world robotic navigation scenarios.*

*Equal contribution.
†Project leader.

1. Introduction

Cross-modal drone navigation represents a key advancement in autonomous robotics, enabling drones to understand and execute navigation instructions derived from natural language descriptions [9, 21–23, 68]. This technology has profound implications for applications ranging from disaster management and search and rescue operations to autonomous surveillance and environmental monitoring [9].

Unlike traditional GPS navigation systems, natural language-guided drone navigation requires a deep understanding of spatial relationships, visual semantics, and cross-modal matching between textual descriptions and aerial imagery [14, 66, 69]. The fundamental challenge in this field

lies in bridging the semantic gap between human-language descriptions and visual representations captured from aerial perspectives. Drone-viewed imagery exhibits unique characteristics, including significant viewpoint variations, scale differences, and complex spatial arrangements that differ significantly from ground-based imagery [2, 3, 34, 41, 42, 45]. These factors render traditional image retrieval methods inadequate for real-world drone navigation applications, where accuracy and reliability are crucial [9, 24, 26, 59].

Existing cross-modal retrieval methods for drone navigation primarily rely on end-to-end learning frameworks that directly map textual queries to visual features. The GeoText-1652 [9] baseline method proposed by Chu et al. represents the current state of the art in this area. It employs a multimodal framework that combines an image encoder (Swin Transformer), a text encoder (BERT) [11], and a cross-modal attention mechanism with spatially aware learning via hybrid spatial matching. Despite these advances, this baseline method still faces inherent limitations in terms of semantic precision and retrieval accuracy. Direct mapping from textual descriptions to visual embeddings often fails to capture subtle semantic relationships, especially when multiple visually similar objects are present in aerial scenes.

To overcome these limitations, we propose ***CGRS (Caption-Guided Retrieval System)***, a novel two-stage retrieval refinement framework that improves semantic precision through intelligent reranking. Specifically, ***CGRS*** operates through a carefully designed two-stage pipeline. In the first stage, we leverage the proven effectiveness of the GeoText-1652 baseline to perform an initial coarse retrieval, obtaining the top 20 most relevant candidate images for each text query. This coarse ranking leverages the baseline's strengths in spatially-aware feature learning and cross-modal alignment, while providing a manageable subset of candidate images for further refinement. The second stage represents our key innovation: we employ a Vision-Language-Model to automatically generate detailed captions for the top 20 candidate images. These generated captions serve as a rich semantic bridge, capturing fine-grained visual details, spatial relationships, and contextual information that may not be fully captured in the original visual embeddings. Subsequently, we perform semantic reranking of the candidate images by computing semantic similarity between the original text query and the captions generated by the VLM. This caption-guided refinement process leverages the descriptive power of natural language to capture subtle visual differences and semantic nuances that direct visual-text matching might miss, resulting in more accurate matching.

Our experimental results on the RoboSense 2025 Track 4 challenge dataset [9] demonstrate the effectiveness of this approach, achieving a consistent 5% improvement across all key metrics (Recall@1, Recall@5, and Recall@10) compared to the baseline method. These improvements ultimately helped us secure a top-2 finish in the competition, validating the practical value of our semantic refinement strategy in real-world cross-modal drone navigation applications. The main contributions are as follows:

- We propose a ***Caption-Guided Retrieval System (CGRS)***, a two-stage coarse-to-fine retrieval framework that leverages visual language models to bridge the semantic gap between natural language queries and aerial imagery.
- We introduce a novel caption-guided reranking mechanism that transforms cross-modal matching into a text-to-text similarity computation, enabling finer-grained semantic alignment in complex aerial scenes.
- Compared to baseline methods, we achieve a consistent 5% improvement on all key metrics (Recall@1, Recall@5, and Recall@10) and win ***TOP-3*** in Cross-Modal Drone Navigation Track of the IROS 2025 RoboSense Challenge.

2. Related Work

2.1. Cross-view Geolocalization

The task of cross-view geolocalization aims to connect visual data captured from distinct viewpoints, such as ground, aerial, or drone perspectives, to their real-world geographic positions. A central difficulty lies in designing representations that remain stable under drastic viewpoint shifts. Early studies attempted to address this by introducing strategies like partitioning images into local regions [57] or adding attention modules that highlight key landmarks [46]. More recent work explores transformer-based designs [10, 63] and hybrid models that integrate local and global features [53]. In parallel, several approaches incorporate auxiliary cues, such as pose estimation [55] or orientation information [27], while others utilize advanced augmentation or contrastive frameworks [12, 72, 75] to improve robustness. Some works also emphasize domain-specific architectures, e.g., cross-drone transformers [4]. In contrast to these directions, our study focuses on new language-driven drone navigation tasks, which naturally bridge human intention and autonomous flight by directly interpreting textual commands.

2.2. Multi-Modality Alignment

Aligning text and visual information has been widely studied in vision-language retrieval [1, 5, 6, 16, 19–23, 25, 49, 50]. Early work emphasized architectural designs such as dual-path frameworks [8, 37, 47, 56, 67, 70, 71, 73, 74]. Later, research shifted to fine-grained alignment: adaptive gating suppressed irrelevant matches [58], and graph-based reasoning captured semantic relations [40]. More recent studies highlight robustness and relational modeling, including uncertainty-aware alignment for cross-domain video-text retrieval [15], multimodal co-attention for video-audio retrieval [18], and attentive relational aggregation for video-text retrieval [17]. With the rise of large-scale pretrain-

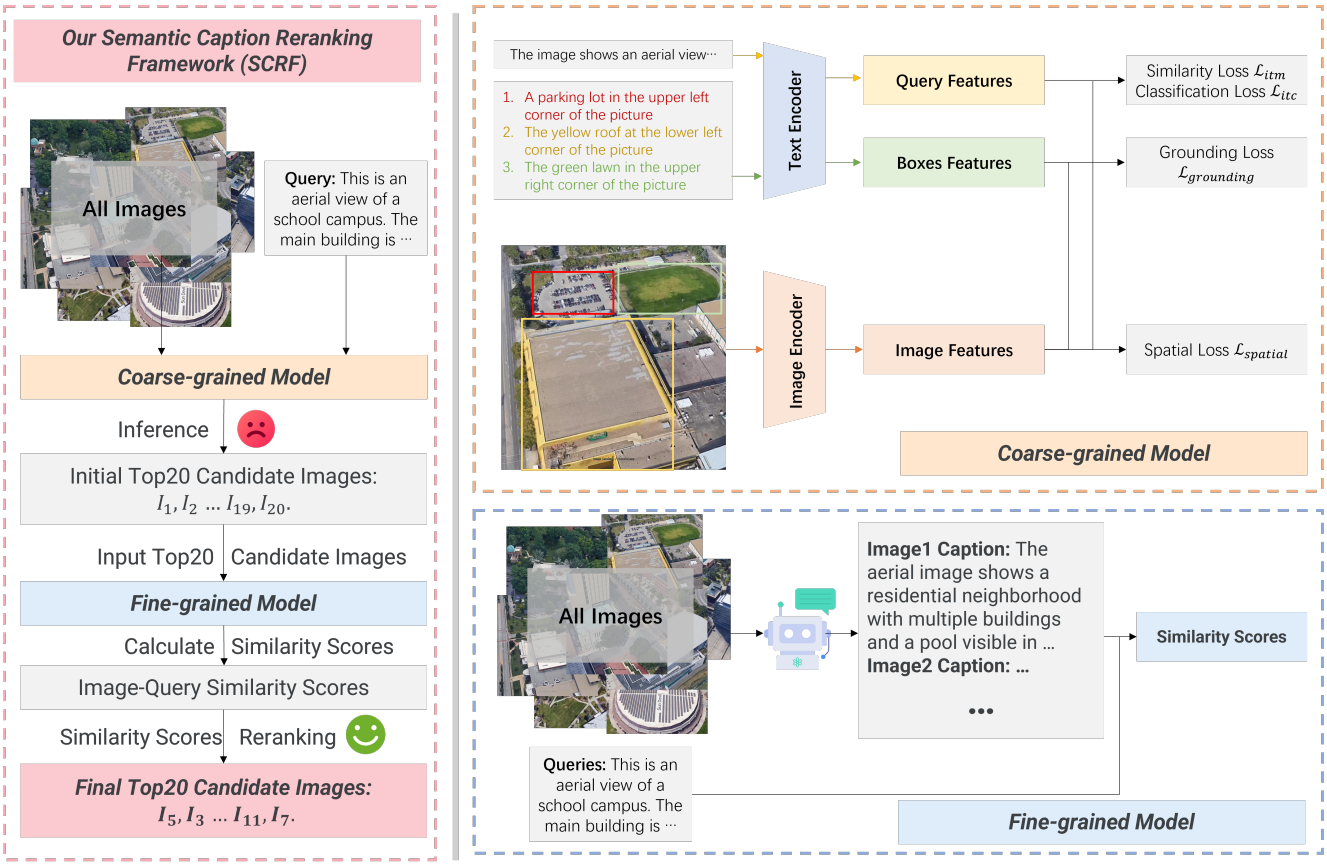

Figure 1. **Overview of our *Caption-Guided Retrieval System (CGRS).*** Our framework employs a two-stage pipeline: the coarse-grained model first retrieves the top 20 candidate images from the gallery using the GeoText-1652 baseline. The fine-grained model then generates detailed captions for these candidates using a Vision-Language Model (VLM) and performs semantic reranking based on text-to-text similarity between the query and generated captions, producing the final top 20 results.

ing, word-region alignment [7], object-tag anchoring [43], and attribute-focused keywords [64] became common practices. At the same time, contrastive pretraining frameworks such as CLIP [52], BLIP [39], and follow-up works [29, 36, 38, 48, 54, 62, 65] achieved remarkable advances in bridging modalities. Unlike these approaches, which largely focus on semantic or global correspondence, our work emphasizes spatial-aware alignment, explicitly modeling fine-grained region-text relations. This design better suits navigation scenarios, where relative positions and local spatial cues are essential for precise drone control.

3. Method

3.1. Problem Definition

The task of natural language-guided drone navigation can be formulated as a text-to-image retrieval problem. Given a natural language query T describing a location with spatial features, the goal is to retrieve the top K most relevant drone-viewed images from a large image library $\mathcal{D} = \{I_1, I_2, ..., I_N\}$.

$$\mathcal{R} = \text{TopK}_{I \in \mathcal{D}}\, s(T, I), \tag{1}$$

where $s : \mathcal{T} \times \mathcal{I} \rightarrow \mathbb{R}$ is a similarity function that measures the semantic and spatial alignment between the text query T and each drone-viewed image I. The output $\mathcal{R} = [I_{\pi(1)}, I_{\pi(2)}, ..., I_{\pi(K)}]$ is a list of images sorted in descending order of similarity score, where π is a permutation such that $s(T, I_{\pi(i)}) \geq s(T, I_{\pi(i+1)})$, where $i \in [1, K-1]$.

The key challenge is to learn a similarity function $s(\cdot, \cdot)$ that captures the fine-grained spatial relationship between textual descriptions and visual regions (e.g., "left", "top right", "center"), thereby enabling the drone to accurately navigate to previously visited locations.

3.2. Method Overview

As shown in Fig. 1, our Caption-Guided Retrieval System (CGRS) employs a two-stage, coarse-to-fine pipeline. In the coarse-grained stage, we use the GeoText-1652 baseline model to retrieve the top 20 candidate images from the

image gallery for each text query and leverage its spatially-aware cross-modal alignment for efficient refinement. In the fine-grained stage, we use the Vision-Language Model (VLM) to generate detailed captions for these candidate images, capturing rich semantic and spatial information. These generated captions serve as a semantic bridge between the visual content and the natural language query. We then compute a similarity score between the original query and each caption and rerank the candidate images based on text-to-text matching (rather than direct visual-text alignment). This caption-guided refinement effectively captures subtle semantic differences that direct visual matching might miss, especially in complex aerial scenes containing multiple similar objects.

3.3. Caption-Guided Retrieval System (CGRS)

Coarse-grained Model Our coarse-grained retrieval stage is built on the GeoText-1652 baseline model [9], which adopts a multimodal framework consisting of an image encoder, a text encoder, and a cross-modal fusion module for spatially aware matching.

Given a text query T and an image I from a gallery $\mathcal{D}$, we use the Swin Transformer as the image encoder to extract visual features $\mathbf{V} \in \mathbb{R}^{d_v}$ and BERT as the text encoder to extract text features $\mathbf{T} \in \mathbb{R}^{d_t}$. For region-level spatial matching, we extract J region features $\{\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_J\}$ from the intermediate feature map via RoI pooling, corresponding to bounding boxes $\{b_1, b_2, ..., b_J\}$, where $b_j = (c_x, c_y, w, h)$.

The model is optimized using multiple complementary loss functions:

We compute the cosine similarity between image and text features as $s(\mathbf{V}, \mathbf{T}) = \frac{\mathbf{V} \cdot \mathbf{T}}{||\mathbf{V}||_2 ||\mathbf{T}||_2}$. The two-way contrastive loss is defined as:

$$\mathcal{L}_{\text{itc}} = -\frac{1}{2}\mathbb{E}\left[\log p_{v2t} + \log p_{t2v}\right], \tag{2}$$

where $p_{v2t} = \frac{\exp(s(\mathbf{V},\mathbf{T})/\tau)}{\sum_{i=1}^{N} \exp(s(\mathbf{V},\mathbf{T}_i)/\tau)}$ and $p_{t2v} = \frac{\exp(s(\mathbf{V},\mathbf{T})/\tau)}{\sum_{i=1}^{N} \exp(s(\mathbf{V}_i,\mathbf{T})/\tau)}$ denote the visual-to-text and text-to-visual similarities, respectively, where τ is a learnable temperature parameter.

The cross-modal encoder predicts whether an image-text pair matches. Hard negative mining is used in each batch, and the binary classification loss is:

$$\mathcal{L}_{\text{itm}} = -\mathbb{E}\left[y_m \log(p_{\text{match}}) + (1 - y_m)\log(1 - p_{\text{match}})\right], \tag{3}$$

where $y_m \in \{0, 1\}$ indicates whether the pair is a positive or negative example.

For each region-level text description $\mathbf{T}_j$, the model predicts a bounding box $\hat{b}_j = (\hat{c}_x, \hat{c}_y, \hat{w}, \hat{h})$ using a criss-cross attention mechanism and a multi-layer perceptron (MLP). The base loss combines ℓ_1 regression and IoU loss:

$$\mathcal{L}_{\text{grounding}} = \mathbb{E}\left[\mathcal{L}_{\text{iou}}(b_j, \hat{b}_j) + ||b_j - \hat{b}_j||_1\right]. \tag{4}$$

To capture relative spatial relationships, we take regional feature pairs $\mathbf{R}_{ij} = [\mathbf{R}_i; \mathbf{R}_j]$ and predict their nine categories of spatial relationships (combining 3 horizontal × 3 vertical positions) through an MLP:

$$\mathcal{L}_{\text{spatial}} = \mathbb{E}\left[-y_{ij}^r \log(\hat{p}_{ij}^r)\right], \tag{5}$$

Where y_{ij}^r is the ground-truth spatial category derived from the bounding box coordinates, and $\hat{p}_{ij}^r$ is the predicted probability distribution.

The total training loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{itm}} + \lambda(\mathcal{L}_{\text{grounding}} + \mathcal{L}_{\text{spatial}}), \tag{6}$$

where $\lambda = 0.1$ balances spatial matching and semantic alignment.

During inference, for each query T, we compute a similarity score $s(T, I_i)$ using the cosine similarity between the global features of all images in the gallery. The formula for selecting the top 20 candidate sets is:

$$\mathcal{C}_{20} = \text{TopK}_{I \in \mathcal{D}, K=20}\, s(T, I). \tag{7}$$

This serves as the input for the subsequent fine-grained reranking stage.

Fine-Grained Model The fine-grained reranking stage leverages a visual language model to generate semantically rich captions for candidate images, enabling more accurate text-to-text similarity matching and capturing subtle semantic relationships.

Caption Generation: For each candidate image $I_k \in \mathcal{C}_{20}$, we employ a pre-trained visual language model (VLM) $\mathcal{M}_{\text{VLM}}$ to generate detailed captions C_k that describe the visual content, spatial layout, and contextual information:

$$C_k = \mathcal{M}_{\text{VLM}}(I_k, \mathcal{P}_{\text{cap}}), \tag{8}$$

where $\mathcal{P}_{\text{cap}}$ is a carefully designed hint template that guides the VLM to generate spatially aware and semantically rich captions. Specifically, we use the following prompts:

```
Please describe this aerial/drone-view
image in detail.  Focus on:  (1)
the main building or structure in
the center of the image and its
architectural features; (2) the
surrounding buildings and their
relative positions (left, right, top,
bottom); (3) significant landmarks
such as parking lots, sports fields,
roads, or vegetation; (4) the overall
spatial layout and arrangement of
objects.  Please be specific and
precise.
```

This prompt explicitly encourages the VLM to capture fine-grained spatial relationships and landmark descriptions,

Table 1. Competition Results on IROS 2025 RoboSense Challenge Cross-Modal Drone Navigation Track. Our team (Xiaomi EV-AD VLA) achieved second place among 8 participating teams. The best results in each metric are shown in **bold**, and our results are highlighted.

Rank	Participant Team	Cross-Modal Drone Navigation Challenge		
		Recall@1 (↑)	Recall@5 (↑)	Recall@10 (↑)
1	lineng	**38.31**	**53.76**	**61.32**
2	**Xiaomi EV-AD VLA (Ours)**	**31.33**	**49.09**	**57.15**
3	rhao_hur	28.34	54.08	**66.11**
4	HiTsz-iLearn	28.21	45.17	52.37
5	RED	26.2	41.4	49.65
6	RoboSense2025	25.44	40.61	49.1
7	testliu	23.14	35.11	41.7
8	geotes	22.54	34.58	41.21

making them consistent with the spatially aware nature of the original text query.

Given a raw text query T and generated captions $\{C_1, C_2, ..., C_{20}\}$, we use a pre-trained sentence embedding model $\mathcal{E}_{\text{sent}}$ to encode the query and captions into a shared semantic space:

$$\mathbf{e}_T = \mathcal{E}_{\text{sent}}(T), \quad \mathbf{e}_{C_k} = \mathcal{E}_{\text{sent}}(C_k), \tag{9}$$

where $\mathbf{e}_T, \mathbf{e}_{C_k} \in \mathbb{R}^{d_e}$ are semantic embeddings. The semantic similarity between the query and each title is calculated as follows:

$$s_{\text{sem}}(T, C_k) = \frac{\mathbf{e}_T \cdot \mathbf{e}_{C_k}}{||\mathbf{e}_T||_2 ||\mathbf{e}_{C_k}||_2}. \tag{10}$$

To simultaneously leverage the visual semantic alignment of the coarse-grained model and the linguistic semantic matching of the captions, we fuse the coarse-grained similarity $s_{\text{coarse}}(T, I_k)$ with the fine-grained semantic similarity $s_{\text{sem}}(T, C_k)$ through a weighted combination:

$$s_{\text{final}}(T, I_k) = \alpha \cdot s_{\text{coarse}}(T, I_k) + (1-\alpha) \cdot s_{\text{sem}}(T, C_k), \tag{11}$$

where $\alpha \in [0, 1]$ is a hyperparameter controlling the balance between visual and semantic matching. In our experiments, we set $\alpha = 0.3$ to emphasize caption-based semantic refinement while preserving the visual alignment signal.

The top 20 candidate words are reranked based on the combined similarity scores, resulting in the final retrieval results:

$$\mathcal{R}_{\text{final}} = \text{Sort}_{\text{desc}}\{(I_k, s_{\text{final}}(T, I_k)) \mid I_k \in \mathcal{C}_{20}\}. \tag{12}$$

This caption-oriented reranking method effectively transforms the challenging cross-modal matching problem into a more tractable text-to-text similarity computation task, where the rich expressiveness of natural language helps bridge subtle semantic gaps that might be overlooked by purely visual features.

4. Experiments

4.1. Benchmark

We use the official data provided by the *RoboSense Challenge 2025* [35] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [31, 32] at ICRA 2023 and the *RoboDrive Challenge 2024* [33, 61] at ICRA 2024, continuing the collective effort to advance robust and scalable robot perception. Each track in this competition is grounded on an established benchmark designed for evaluating real-world robustness and generalization [9, 13, 30, 44, 45, 60]. Specifically, this task is built upon the **GeoText-1652** dataset [9] in **Track 4**, which benchmarks cross-modal image-text retrieval for language-guided drone navigation across drastically different viewpoints and real-world sensing conditions.

The GeoText-1652 dataset [9] is a large-scale multimodal benchmark designed for natural language-guided drone geolocalization. The training set contains 50,218 images from 33 universities, with 113,562 global descriptions and 113,367 region-level bounding box-text pairs; the test set contains 54,227 images from 39 non-overlapping universities, with 154,065 global descriptions and 140,179 bounding box-text pairs. Each image is annotated with an average of 3 global descriptions (70.23 words each) and 2.62 region-level descriptions (21.6 words each), which explicitly capture spatial relationships such as "left", "top right", and "center". This fine-grained spatial annotation makes GeoText-1652 [9] particularly challenging and suitable for evaluating cross-modal retrieval methods that require precise spatial understanding.

4.2. Evaluation Metrics

Following the standard protocol established by the benchmark and challenge [9], we evaluated our approach on the multimodal drone navigation task, which employs text-to-image retrieval to retrieve relevant drone-viewed images

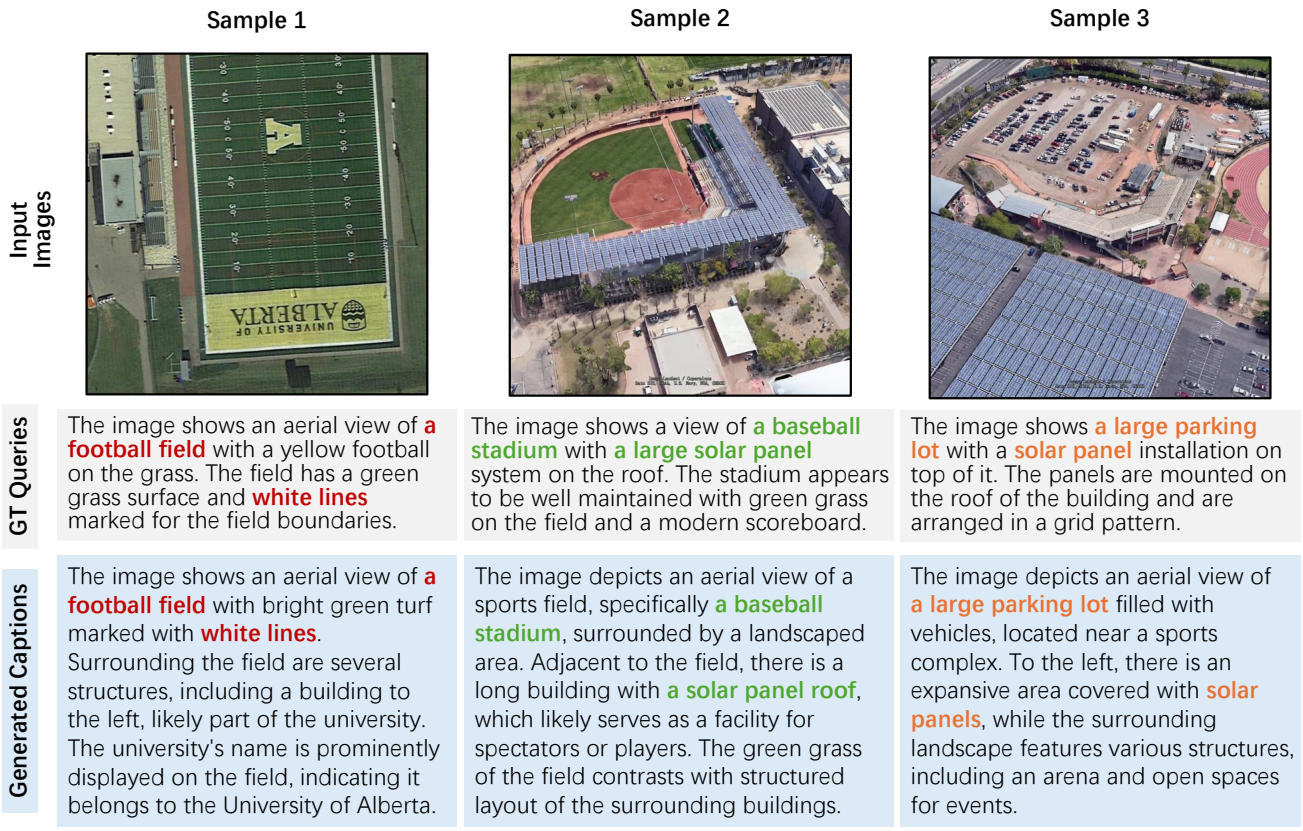

Figure 2. **Qualitative Results.**

using natural language queries. We used Recall@K (R@K) as the primary evaluation metric, measuring the percentage of queries for which at least one correct image appears in the top K retrieval results. Specifically, we report R@1, R@5, and R@10 to assess the accuracy of fine-grained matching and the robustness of the overall retrieval. Higher Recall@K indicates better retrieval performance, while R@1 reflects the model's ability to prioritize correct images.

4.3. Implementation Details

We replicated the GeoText-1652 baseline model [9] on 8 NVIDIA A800 GPUs for 5 epochs, taking approximately 30 GPU hours. We used the AdamW [51] optimizer with a learning rate of 3×10^{-5} and a spatial matching weight $\lambda = 0.1$. After training, we used this baseline model to extract the top 20 candidate images for each test query. In the fine-grained reranking stage, we used GPT-4o [28] with a maximum token count of 256 to generate detailed captions for all candidate images. The average length of the generated captions was 120-150 words, capturing spatial layout and landmark details. We then used BERT [11] as the sentence encoder $\mathcal{E}_{\text{sent}}$ to compute the semantic similarity between the original query and the generated captions. After validation experiments, the hybrid fusion weight was empirically set to $\alpha = 0.3$, prioritizing semantic matching based on captions while preserving the visual signals from the coarse model. The entire caption generation process takes approximately 12 hours of offline time, while the final reranking process adds almost no inference overhead (¡ 10 milliseconds per query).

4.4. Challenge Results

Table 1 shows the official leaderboard for Track 4 of the IROS 2025 RoboSense Challenge. Our method stands out among eight participating teams, ranking second with R@1 scores of 31.33%, R@5 scores of 49.09%, and R@10 scores of 57.15%. Our R@1 score is 5.89% higher than the official baseline, "RoboSense2025", 8.48% higher in R@5, and 8.05% higher in R@10, validating the effectiveness of our caption-guided semantic refinement strategy. While the winning team "lineng" achieved an R@1 that was 6.98% higher than the third-place team "rhao_hur", our approach exhibited more balanced performance compared to the third-place team's R@1 (28.34% R@1 vs. 66.11% R@10), maintaining high accuracy at the top of the rankings, which is crucial for practical drone navigation, as the top-ranked predictions directly determine the target location. Our approach consistently improved compared to the lower-ranked teams

(4th to 8th place, with R@1 values ranging from 22.54% to 28.21%), highlighting that leveraging visual language models for explicit caption generation provides richer semantic representations than direct visual-text alignment, and is particularly beneficial for capturing fine-grained spatial relationships in drone-viewed imagery.

4.5. Qualitative Analysis

Figure 2 shows the quality of caption generation for three representative samples. VLM is able to accurately capture fine-grained spatial details and semantic relationships in drone-view imagery. In sample 1, it correctly identifies the football field with yellow markings and white boundary lines, and is even able to identify the University of Alberta from visible text. Sample 2 demonstrates the model's ability to describe complex layouts, capturing the baseball field, the adjacent solar panel roof, and the surrounding landscape area. Sample 3 excels at distinguishing similar structures—the generated caption distinguishes a large parking lot with solar panels from a nearby sports facility, providing rich contextual information that is critical for matching queries with refined spatial descriptors. These examples verify that VLM-generated captions are able to effectively convert visual content into linguistically rich descriptions that naturally align with human-written queries, thereby facilitating more reliable semantic matching in the reranking stage.

5. Conclusion

This paper proposes a caption-guided retrieval system that addresses the challenge of natural language-guided drone navigation through a two-stage, coarse-to-fine retrieval pipeline. First, we leverage a baseline for efficient coarse-grained filtering. Then, we leverage a visual language model to generate semantically rich captions and perform fine-grained reranking, effectively bridging the semantic gap between text queries and aerial imagery. Compared to the baseline, our approach achieves consistent improvements of approximately 5-8% across all recall metrics and achieved a ***Top-2*** ranking among eight participating teams in the IROS 2025 RoboSense Challenge Track4. These results demonstrate the practical value of leveraging large-scale visual language models for semantic refinement in real-world robotic navigation scenarios, demonstrating that explicit language representations can serve as an effective medium for cross-modal understanding in spatially complex aerial environments.

References

[1] Samuel Albanie, Yang Liu, Arsha Nagrani, Antoine Miech, Ernesto Coto, Ivan Laptev, Rahul Sukthankar, Bernard Ghanem, Andrew Zisserman, Valentin Gabeur, et al. The end-of-end-to-end: A video understanding pentathlon challenge (2020). *arXiv preprint arXiv:2008.00744*, 2020.

[2] Hengwei Bian et al. Dynamiccity: Large-scale 4D occupancy generation from dynamic scenes. In *International Conference on Learning Representations*, 2025.

[3] Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J Taylor, and Kostas Daniilidis. M3ED: Multi-robot, multi-sensor, multi-environment event dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4016–4023, 2023.

[4] Guanlin Chen, Pengfei Zhu, Bing Cao, Xing Wang, and Qinghua Hu. Cross-drone transformer network for robust single object tracking. *IEEE transactions on circuits and systems for video technology*, 33(9):4552–4563, 2023.

[5] Runnan Chen et al. Clip2scene: Towards label-efficient 3D scene understanding by CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[6] Runnan Chen et al. Towards label-free scene understanding by vision foundation models. In *Advances in Neural Information Processing Systems*, pages 75896–75910, 2023.

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[8] Hao Cheng, Erjia Xiao, Yichi Wang, Lingfeng Zhang, Qiang Zhang, Jiahang Cao, Kaidi Xu, Mengshu Sun, Xiaoshuai Hao, Jindong Gu, et al. Exploring typographic visual prompts injection threats in cross-modality generation models. *arXiv preprint arXiv:2503.11519*, 2025.

[9] Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, pages 213–231, 2024.

[10] Ming Dai, Jianhong Hu, Jiedong Zhuang, and Enhui Zheng. A transformer-based feature segmentation and region alignment method for uav-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4376–4389, 2021.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Aayush Dhakal, Adeel Ahmad, Subash Khanal, Srikumar Sastry, Hannah Kerner, and Nathan Jacobs. Sat2cap: Mapping fine-grained textual descriptions from satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 533–542, 2024.

[13] Zeying Gong, Tianshuai Hu, Ronghe Qiu, and Junwei Liang. From cognition to precognition: A future-aware framework for social navigation. In *IEEE International Conference on Robotics and Automation*, pages 9122–9129, 2025.

[14] Zeying Gong, Rong Li, Tianshuai Hu, Ronghe Qiu, Lingdong Kong, Lingfeng Zhang, Yiyi Ding, Leying Zhang, and Junwei Liang. Stairway to success: Zero-shot floor-aware object-goal navigation via llm-driven coarse-to-fine exploration. *arXiv preprint arXiv:2505.23019*, 2025.

[15] Xiaoshuai Hao and Wanqian Zhang. Uncertainty-aware alignment network for cross-domain video-text retrieval. *Advances in Neural Information Processing Systems*, 36:38284–38296, 2023.

[16] Xiaoshuai Hao, Yucan Zhou, Dayan Wu, Wanqian Zhang, Bo Li, and Weiping Wang. Multi-feature graph attention network for cross-modal video-text retrieval. In *Proceedings of the 2021 international conference on multimedia retrieval*, pages 135–143, 2021.

[17] Xiaoshuai Hao, Yucan Zhou, Dayan Wu, Wanqian Zhang, Bo Li, Weiping Wang, and Dan Meng. What matters: Attentive and relational feature aggregation network for video-text retrieval. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE Computer Society, 2021.

[18] Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. Listen and look: Multi-modal aggregation and co-attention network for video-audio retrieval. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[19] Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. Dual alignment unsupervised domain adaptation for video-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18962–18972, 2023.

[20] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. Mixgen: A new multi-modal data augmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 379–389, 2023.

[21] Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, Haimei Zhao, Hui Zhang, Yi Zhou, Qiang Wang, Weiming Li, Lingdong Kong, and Jing Zhang. Is your hd map constructor reliable under sensor corruptions? *Advances in Neural Information Processing Systems*, 37:22441–22482, 2024.

[22] Xiaoshuai Hao, Hui Zhang, Yifan Yang, Yi Zhou, Sangil Jung, Seung-In Park, and ByungIn Yoo. Mbfusion: A new multi-modal bev feature fusion method for hd map construction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15922–15928. IEEE, 2024.

[23] Xiaoshuai Hao, Yunfeng Diao, Mengchuan Wei, Yifan Yang, Peng Hao, Rong Yin, Hui Zhang, Weiming Li, Shu Zhao, and Yu Liu. Mapfusion: A novel bev feature fusion network for multi-modal map construction. *Information Fusion*, 119: 103018, 2025.

[24] Xiaoshuai Hao, Guanqun Liu, Yuting Zhao, et al. Msc-bench: Benchmarking and analyzing multi-sensor corruption for driving perception. *arXiv preprint arXiv:2501.01037*, 2025.

[25] Xiaoshuai Hao, Haimei Zhao, Yunfeng Diao, Rong Yin, Guangyin Jin, Jing Zhang, Wanqian Zhang, and Wei Zhou. Dada++: Dual alignment domain adaptation for unsupervised video-text retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

[26] Xiaoshuai Hao et al. Safemap: Robust HD map construction from incomplete observations. In *International Conference on Machine Learning*, pages 22091–22102. PMLR, 2025.

[27] Wenmiao Hu, Yichen Zhang, Yuxuan Liang, Yifang Yin, Andrei Georgescu, An Tran, Hannes Kruppa, See-Kiong Ng, and Roger Zimmermann. Beyond geo-localization: Fine-grained orientation of street-view images by cross-view matching with satellite imagery. In *Proceedings of the 30th ACM international conference on multimedia*, pages 6155–6164, 2022.

[28] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[29] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[30] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3D: Towards robust and reliable 3D perception against corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023.

[31] Lingdong Kong, Yaru Niu, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, Liangjun Zhang, Hesheng Wang, Wei Tsang Ooi, Ruijie Zhu, Ziyang Song, Li Liu, Tianzhu Zhang, Jun Yu, Mohan Jing, Pengwei Li, Xiaohua Qi, Cheng Jin, Yingfeng Chen, Jie Hou, Jie Zhang, Zhen Kan, Qiang Lin, Liang Peng, Minglei Li, Di Xu, Changpeng Yang, Yuanqi Yao, Gang Wu, Jian Kuai, Xianming Liu, Junjun Jiang, Jiamian Huang, Baojun Li, Jiale Chen, Shuang Zhang, Sun Ao, Zhenyu Li, Runze Chen, Haiyong Luo, Fang Zhao, and Jingze Yu. The RoboDepth challenge: Methods and advancements towards robust depth estimation. *arXiv preprint arXiv:2307.15061*, 2023.

[32] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit R. Cottereau, and Wei Tsang Ooi. RoboDepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems*, pages 21298–21342, 2023.

[33] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottereau, Lai Xing Ng, Yuexin Ma, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, Weichao Qiu, Wei Zhang, Xu Cao, Hao Lu, Ying-Cong Chen, Caixin Kang, Xinning Zhou, Chengyang Ying, Wentao Shang, Xingxing Wei, Yinpeng Dong, Bo Yang, Shengyin Jiang, Zeliang Ma, Dengyi Ji, Haiwen Li, Xingliang Huang, Yu Tian, Genghua Kou, Fan Jia, Yingfei Liu, Tiancai Wang, Ying Li, Xiaoshuai Hao, Yifan Yang, Hui Zhang, Mengchuan Wei, Yi Zhou, Haimei Zhao, Jing Zhang, Jinke Li, Xiao He, Xiaoqiang Cheng, Bingyang Zhang, Lirong Zhao, Dianlei Ding, Fangsheng Liu, Yixiang Yan, Hongming Wang, Nanfei Ye, Lun Luo, Yubo Tian, Yiwei Zuo, Zhe Cao, Yi Ren, Yunfan Li, Wenjie Liu, Xun Wu, Yifan Mao, Ming Li, Jian Liu, Jiayang Liu, Zihan Qin, Cunxi Chu, Jialei Xu, Wenbo Zhao, Junjun Jiang, Xianming Liu, Ziyan Wang, Chiwei Li, Shilong Li, Chendong Yuan, Songyue Yang, Wentao Liu, Peng Chen, Bin Zhou, Yubo Wang, Chi Zhang, Jianhang Sun, Hai Chen, Xiao Yang, Lizhong Wang, Dongyi Fu, Yongchun Lin, Huitong Yang, Haoang Li, Yadan Luo, Xianjing Cheng, and Yong Xu. The RoboDrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.

[34] Lingdong Kong, Dongyue Lu, Xiang Xu, Lai Xing Ng, Wei Tsang Ooi, and Benoit R. Cottereau. Eventfly: Event camera perception from ground to the sky. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1472–1484, 2025.

[35] Lingdong Kong, Shaoyuan Xie, Zeying Gong, Ye Li, Meng Chu, Ao Liang, Yuhao Dong, Tianshuai Hu, Ronghe Qiu, Rong Li, Hanjiang Hu, Dongyue Lu, Wei Yin, Wenhao Ding, Linfeng Li, Hang Song, Wenwei Zhang, Yuexin Ma, Junwei Liang, Zhedong Zheng, Lai Xing Ng, Benoit R. Cottereau, Wei Tsang Ooi, Ziwei Liu, Zhanpeng Zhang, Weichao Qiu, Wei Zhang, Ji Ao, Jiangpeng Zheng, Siyu Wang, Guang Yang, Zihao Zhang, Yu Zhong, Enzhu Gao, Xinhan Zheng, Xueting Wang, Shouming Li, Yunkai Gao, Siming Lan, Mingfei Han, Xing Hu, Dusan Malic, Christian Fruhwirth-Reisinger, Alexander Prutsch, Wei Lin, Samuel Schulter, Horst Possegger, Linfeng Li, Jian Zhao, Zepeng Yang, Yuhang Song, Bojun Lin, Tianle Zhang, Yuchen Yuan, Chi Zhang, Xuelong Li, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, Aodi Wu, Xubo Luo, Erjia Xiao, Lingfeng Zhang, Yingbo Tang, Hao Cheng, Renjing Xu, Wenbo Ding, Lei Zhou, Long Chen, Hangjun Ye, Xiaoshuai Hao, Shuangzhi Li, Junlong Shen, Xingyu Li, Hao Ruan, Jinliang Lin, Zhiming Luo, Yu Zang, Cheng Wang, Hanshi Wang, Xijie Gong, Yixiang Yang, Qianli Ma, Zhipeng Zhang, Wenxiang Shi, Jingmeng Zhou, Weijun Zeng, Kexin Xu, Yuchen Zhang, Haoxiang Fu, Ruibin Hu, Yanbiao Ma, Xiyan Feng, Wenbo Zhang, Lu Zhang, Yunzhi Zhuge, Huchuan Lu, You He, Seungjun Yu, Junsung Park, Youngsun Lim, Hyunjung Shim, Faduo Liang, Zihang Wang, Yiming Peng, Guanyu Zong, Xu Li, Binghao Wang, Hao Wei, Yongxin Ma, Yunke Shi, Shuaipeng Liu, Dong Kong, Yongchun Lin, Huitong Yang, Liang Lei, Haoang Li, Xinliang Zhang, Zhiyong Wang, Xiaofeng Wang, Yuxia Fu, Yadan Luo, Djamahl Etchegaray, Yang Li, Congfei Li, Yuxiang Sun, Wenkai Zhu, Wang Xu, Linru Li, Longjie Liao, Jun Yan, Benwu Wang, Xueliang Ren, Xiaoyu Yue, Jixian Zheng, Jinfeng Wu, Shurui Qin, Wei Cong, and Yao He. The RoboSense challenge: Sense anything, navigate anywhere, adapt across platforms. https://robosense2025.github.io, 2025.

[36] Lingdong Kong, Xiang Xu, Youquan Liu, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. LargeAD: Large-scale cross-sensor data pretraining for autonomous driving. *arXiv preprint arXiv:2501.04005*, 2025.

[37] Dasong Li, Sizhuo Ma, Hang Hua, Wenjie Li, Jian Wang, Chris Wei Zhou, Fengbin Guan, Xin Li, Zihao Yu, Yiting Lu, et al. Vquala 2025 challenge on engagement prediction for short videos: Methods and results. *arXiv preprint arXiv:2509.02969*, 2025.

[38] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[39] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[40] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4654–4662, 2019.

[41] Rong Li, Yuhao Dong, Tianshuai Hu, Ao Liang, et al. 3eed: Ground everything everywhere in 3D. *arXiv preprint arXiv:2511.01755*, 2025.

[42] Rong Li et al. Seeground: See and ground for zero-shot open-vocabulary 3D visual grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3707–3717, 2025.

[43] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision*, pages 121–137. Springer, 2020.

[44] Ye Li, Lingdong Kong, Hanjiang Hu, Xiaohao Xu, and Xiaonan Huang. Is your LiDAR placement optimized for 3D scene understanding? In *Advances in Neural Information Processing Systems*, pages 34980–35017, 2024.

[45] Ao Liang et al. Perspective-invariant 3D object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 27725–27738, 2025.

[46] Jinliang Lin, Zhedong Zheng, Zhun Zhong, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing*, 31:3780–3792, 2022.

[47] Peiran Liu, Qiang Zhang, Daojie Peng, Lingfeng Zhang, Yihao Qin, Hang Zhou, Jun Ma, Renjing Xu, and Yiding Ji. Toponav: Topological graphs as a key enabler for advanced object navigation. *arXiv preprint arXiv:2509.01364*, 2025.

[48] Youquan Liu et al. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, pages 37193–37229, 2023.

[49] Youquan Liu et al. Uniseg: A unified multi-modal LiDAR segmentation network and the openpcseg codebase. In *IEEE/CVF International Conference on Computer Vision*, pages 21662–21673, 2023.

[50] Youquan Liu et al. Multi-space alignments towards universal lidar segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14648–14661, 2024.

[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[53] Royston Rodrigues and Masahiro Tani. Global assists local: Effective aerial representations for field of view constrained image geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3871–3879, 2022.

[54] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9891–9901, 2022.

[55] Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17010–17020, 2022.

[56] BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, Yi Han, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025.

[57] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):867–879, 2021.

[58] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5764–5773, 2019.

[59] Yujie Wu, Huaihai Lyu, Yingbo Tang, Lingfeng Zhang, Zhihui Zhang, Wei Zhou, and Siqi Hao. Evaluating gpt-4o's embodied intelligence: A comprehensive empirical study. *Authorea Preprints*, 2025.

[60] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are VLMs ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. In *IEEE/CVF International Conference on Computer Vision*, pages 6585–6597, 2025.

[61] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3878–3894, 2025.

[62] Xiang Xu et al. Beyond one shot, beyond one perspective: Cross-view and long-horizon distillation for better lidar representations. In *IEEE/CVF International Conference on Computer Vision*, pages 25506–25518, 2025.

[63] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34:29009–29020, 2021.

[64] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM international conference on multimedia*, pages 4492–4501, 2023.

[65] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

[66] Lingfeng Zhang, Qiang Zhang, Hao Wang, Erjia Xiao, Zixuan Jiang, Honglei Chen, and Renjing Xu. Trihelper: Zero-shot object navigation with dynamic assistance. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 10035–10042, 2024.

[67] Lingfeng Zhang, Xiaoshuai Hao, Yingbo Tang, Haoxiang Fu, Xinyu Zheng, Pengwei Wang, Zhongyuan Wang, Wenbo Ding, and Shanghang Zhang. $nava^3$: Understanding any instruction, navigating anywhere, finding anything. *arXiv preprint arXiv:2508.04598*, 2025.

[68] Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. In *The 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

[69] Lingfeng Zhang, Hao Wang, Erjia Xiao, Xinyao Zhang, Qiang Zhang, Zixuan Jiang, and Renjing Xu. Multi-floor zero-shot object navigation policy. In *IEEE International Conference on Robotics and Automation*, pages 6416–6422, 2025.

[70] Qiang Zhang, Zhang Zhang, Wei Cui, Jingkai Sun, Jiahang Cao, Yijie Guo, Gang Han, Wen Zhao, Jiaxu Wang, Chenghao Sun, et al. Humanoidpano: Hybrid spherical panoramic-lidar cross-modal perception for humanoid robots. *arXiv preprint arXiv:2503.09010*, 2025.

[71] Shuyi Zhang, Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Pengwei Wang, Zhongyuan Wang, Hongxuan Ma, and Shanghang Zhang. Video-cot: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought. *arXiv preprint arXiv:2506.08817*, 2025.

[72] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3480–3488, 2023.

[73] Xinyu Zheng, Yangfan He, Yuhao Luo, Lingfeng Zhang, Jianhui Wang, Tianyu Shi, and Yun Bai. Railway side slope hazard detection system based on generative models. *IEEE Sensors Journal*, 2025.

[74] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.

[75] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022.