A Parameter-Efficient Mixture-of-Experts Framework for Cross-Modal Geo-Localization

LinFeng Li[1,2,*]   Jian Zhao[1,*,†]   Zepeng Yang[1]   Yuhang Song[3]   Bojun Lin[3]
Tianle Zhang[1,†]   Yuchen Yuan[1,†]   Chi Zhang[1,†]   Xuelong Li[1,†]

[1]The Institute of Artificial Intelligence (TeleAI), China Telecom
[2]East China Normal University
[3]National Tsing Hua University

Abstract

We present a winning solution to RoboSense 2025 Track 4: Cross-Modal Drone Navigation. The task retrieves the most relevant geo-referenced image from a large multi-platform corpus (satellite/drone/ground) given a natural-language query. Two obstacles are severe inter-platform heterogeneity and a domain gap between generic training descriptions and platform-specific test queries. We mitigate these with a domain-aligned preprocessing pipeline and a Mixture-of-Experts (MoE) framework: (i) platform-wise partitioning, satellite augmentation, and removal of orientation words; (ii) an LLM-based caption refinement pipeline to align textual semantics with the distinct visual characteristics of each platform. Using BGE-M3 (text) and EVA-CLIP (image), we train three platform experts using a progressive two-stage, hard-negative mining strategy to enhance discriminative power, and fuse their scores at inference. The system tops the official leaderboard, demonstrating robust cross-modal geo-localization under heterogeneous viewpoints.

1. Introduction

Cross-modal geo-localization, which aims to retrieve geo-referenced images from heterogeneous sources given natural language or visual queries, has emerged as a fundamental capability for autonomous navigation, situational awareness, and emergency response [1–6]. In particular, unmanned aerial vehicles (UAVs) play an increasingly critical role in tasks such as disaster management, infrastructure inspection, and urban planning, where robust geo-localization enables accurate scene understanding under diverse viewpoints [7–10]. However, building a generalizable model for cross-modal retrieval across drastically different platforms—satellite, drone, and ground-level imagery—remains highly challenging.

Two key obstacles hinder progress in this domain. First, the data heterogeneity across platforms introduces severe appearance gaps: satellite imagery exhibits large-scale, top-down structures, drone imagery captures mid-level oblique views, while ground-view images contain rich local details with clutter and occlusion [11, 12]. These discrepancies render a single, unified model less effective. Second, a significant domain gap exists between training and evaluation texts: training captions are often generic or verbose, whereas test queries are concise and intent-driven. More critically, the semantic focus of the descriptions often mismatches the visual modality (*e.g.*, a generic caption may fail to capture the specific details relevant to a satellite or drone perspective), leading to poor generalization.

Existing approaches in vision-language retrieval typically rely on large pre-trained encoders such as CLIP [13–22] or ALIGN [23–25] to learn a shared embedding space. While effective on in-domain benchmarks, these methods often struggle to reconcile heterogeneous views and distributional discrepancies without costly fine-tuning on massive curated datasets. Ensemble strategies and Mixture-of-Experts (MoE) [26, 27] methods offer a promising direction by combining specialized models, but most existing designs incur high parameter overhead or lack mechanisms to bridge textual domain gaps.

To address these challenges, we propose the Parameter-Efficient Mixture-of-Experts (PE-MoE) framework, a divide-and-conquer solution that integrates domain-aligned preprocessing with a lightweight expert design. Our framework partitions the dataset by platform, enabling each expert to specialize in satellite, drone, or ground imagery, while sharing a frozen backbone of strong pre-trained encoders (BGE-M3 [28] for text, EVA-CLIP [29] for images) to preserve

generalization. To reduce the textual domain gap, we introduce an LLM-based caption refinement strategy. This process automatically revises captions to ensure their semantic focus aligns with the visual modality (*e.g.*, emphasizing spatial relations for satellite images vs. object details for drone images), creating more precise training pairs. For satellite imagery, we further apply targeted augmentations alongside directional-text sanitization to ensure semantic consistency. The experts are trained using a progressive two-stage, hard-negative mining strategy to sharpen their discriminative abilities. Finally, a dynamic gating network adaptively routes queries to the most relevant experts, producing a fused similarity score.

This design achieves robust retrieval under severe viewpoint and modality shifts while maintaining parameter efficiency. On the RoboSense 2025 Track 4: Cross-Modal Drone Navigation, our method ranked first on the official leaderboard, demonstrating superior performance and strong generalization. Beyond competition success, our study highlights the importance of jointly addressing data heterogeneity and domain alignment, opening new directions for efficient cross-modal geo-localization.

2. Related Work

2.1. Cross-Modal Image–Text Retrieval

Cross-modal retrieval aims to map images and texts into a shared embedding space where their semantic similarity can be effectively measured. Early methods typically combined recurrent neural networks for text encoding with CNN-based visual backbones, optimized using ranking or triplet losses. With the introduction of large-scale vision–language pre-training, models such as CLIP [30–40], ALIGN [41–45], and BLIP [46–49] substantially improved retrieval performance by leveraging millions of image–text pairs. More recent systems like BLIP-2 [50–54] adopt parameter-efficient strategies that freeze pretrained encoders and incorporate lightweight adapters for downstream alignment. Despite their success, most of these models assume relatively homogeneous data domains, where visual content or camera viewpoints remain within a limited distribution. When applied to UAV-based geo-localization, where text queries must be matched against highly heterogeneous satellite, drone, and ground-view images, their performance drops sharply due to the severe modality and viewpoint shifts. This motivates the need for domain-aware or platform-specialized retrieval mechanisms.

2.2. Visual Geo-Localization

Visual geo-localization focuses on determining the geographic location of an input image by retrieving matching reference imagery. Traditional approaches rely on hand-crafted local features [55, 56] and structure-based retrieval pipelines [57, 58], which generally fail under large-scale viewpoint, altitude, or environmental changes. Deep learning methods have significantly pushed the field forward, especially in cross-view matching tasks that bridge ground and aerial imagery [59–61]. Benchmark datasets such as CVUSA [62] and University-1652 [11] have demonstrated the potential of cross-view learning, yet they also reveal the difficulty of aligning visual content captured from drastically different perspectives. In practice, UAV-based retrieval systems face even greater challenges due to the heterogeneity of sensing platforms and the mismatch between verbose training captions and concise test queries. This discrepancy underscores the importance of both visual domain adaptation and linguistic domain alignment for robust UAV geo-localization.

2.3. Mixture-of-Experts and Model Ensembles

Model ensembles and Mixture-of-Experts (MoE) frameworks have long been used to improve accuracy and robustness by combining multiple specialized learners. Traditional ensembles operate by aggregating the predictions of independently trained models, while MoE architectures explicitly introduce expert networks and learn a gating mechanism to route inputs adaptively [53, 55, 63–65]. Recent parameter-efficient MoE designs further integrate frozen backbones with compact expert modules, offering strong specialization without significant computational overhead. In multimodal learning, MoE methods have been applied to vision–language pre-training [58, 66, 67], enabling models to better handle diverse image or text distributions. However, their application to UAV cross-modal geo-localization remains relatively unexplored. Our work advances this direction by designing a parameter-efficient MoE architecture jointly with domain-aligned preprocessing, allowing experts to specialize in platform-specific imagery while maintaining strong generalization across modality and viewpoint gaps.

3. Methodology

In this section, we elaborate on the technical framework of our proposed solution, the Parameter-Efficient Mixture-of-Experts (PE-MoE). Our core philosophy follows a "divide and conquer" principle, aiming to efficiently address the challenges of data heterogeneity and domain gaps by sharing generalized knowledge while specializing in specific domains. As illustrated in Figure 1, our framework is comprised of three primary stages: data preprocessing and alignment, the PE-MoE model architecture, and a two-stage training strategy.

3.1. Data Preprocessing and Alignment

We posit that targeted data preprocessing is a critical prerequisite for model success. Our strategy focuses on stratifying
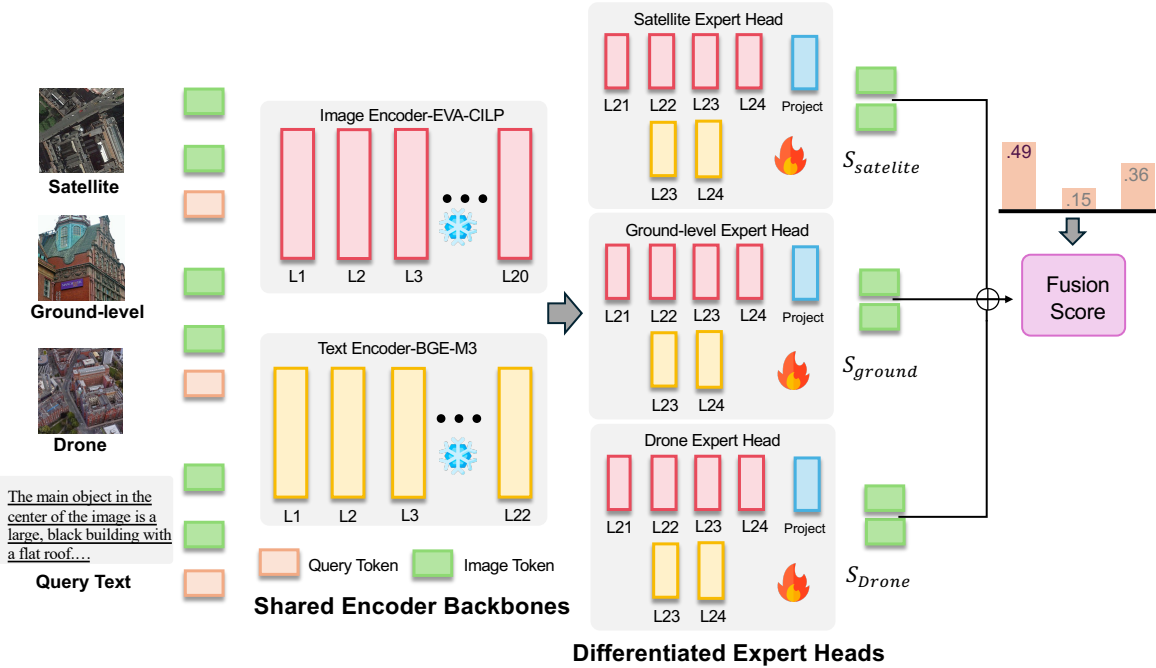
Figure 1. The overall architecture of our Parameter-Efficient Mixture-of-Experts (PE-MoE) framework. A shared backbone extracts general features, which are processed by a dynamic gating network and specialized expert heads to produce the final retrieval score.

data by domain and aligning the textual distributions between training and testing phases.

Platform-based Data Stratification To tackle the profound visual discrepancies across platforms, we first partition the entire training dataset, D, into three distinct, non-overlapping subsets based on the image source: a satellite imagery subset, D_{sat}; a drone imagery subset, D_{drone}; and a ground-view imagery subset, D_{ground}. This stratification allows us to train highly specialized expert models for each visual domain.

Textual Domain Alignment We identified a significant domain gap in the textual descriptions relative to their corresponding image modalities. For example, the focus of a caption for a satellite image should differ substantially from that of a drone-view image (e.g., broad area relations vs. specific object details). To address this, we employed an LLM-based Caption Refinement strategy. We utilized a Large Language Model (LLM) to review and revise the caption for each training image. This process ensured that the textual description was semantically aligned with the image's specific visual perspective (satellite, drone, or ground). By tailoring the captions to be domain-specific, we provide the model with more accurate and consistent text-image pairs, enhancing the specialization of each expert.

Augmentation and Sanitization for Satellite Imagery Given the relatively small sample size of the satellite subset D_{sat}, we applied a series of data augmentation techniques, including random geometric transformations (e.g., rotations, flips) and photometric adjustments (e.g., brightness, contrast jitter). However, geometric transformations can alter the absolute spatial orientation of an image, creating semantic inconsistencies with textual descriptions containing directional language (e.g., "to the north of," "on the left side"). To resolve this, we employed a complementary text sanitization process. Before applying geometric augmentations, a keyword-matching algorithm automatically removed any sentences with explicit directional phrases from the corresponding captions, ensuring semantic consistency between the augmented images and their textual descriptions.

3.2. Parameter-Efficient MoE Framework

Our model architecture is designed to achieve maximum specialization with minimal parameter overhead.

Shared Encoder Backbones We utilize the state-of-the-art BGE-M3 [28] as our text encoder and EVA-CLIP [68] as our image encoder. To maximize parameter efficiency and preserve their powerful, general-purpose representational abilities, the vast majority of the parameters in these backbone models are **kept frozen** during training. Any input text or image undergoes a single forward pass through these

shared backbones to yield high-level, generalized feature representations, denoted as t_{shared} and $v_{\text{raw\_shared}}$.

Differentiated Expert Heads Building upon the shared backbones, we designed three lightweight expert heads, one for each platform: H_{sat}, H_{drone}, and H_{ground}. Each expert head is an independent, trainable module comprising:
- The final few (e.g., 2) trainable transformer layers of the BGE-M3 and EVA-CLIP models.
- A distinct, trainable visual projection layer that maps image features into the common embedding space.

Each expert head H_k is trained exclusively on its corresponding data subset D_k. It takes the shared features as input and processes them to generate domain-specific final embeddings (t_k, v_k), from which a similarity score $S_k(q, I) = \text{cosine}(t_k, v_k)$ is computed.

Dynamic Gating Network To intelligently orchestrate the experts, we designed a dynamic gating network, G. It is a small, two-layer Multi-Layer Perceptron (MLP) that takes the shared text feature t_{shared} as input. Its output is a 3-dimensional logits vector, which is passed through a Softmax function to produce a query-dependent weight distribution $g(q) = [g_{sat}, g_{drone}, g_{ground}]$, where $\sum_k g_k(q) = 1$. The gate learns to "understand" the query's intent and assign the highest weight to the expert best suited to handle it.

3.3. Training and Inference

Two-Stage Training Strategy As illustrated in Figure 2, our training follows a progressive two-stage strategy.
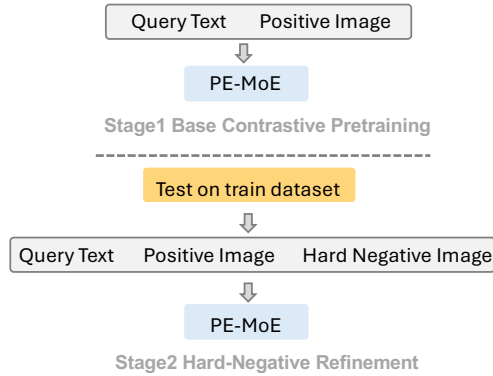

Figure 2. The two-stage training pipeline. Stage 1 builds general alignment using positive pairs, while Stage 2 uses mined hard negatives to refine the model's discriminative ability.

In Stage 1 (Base Contrastive Pretraining), we train the PE-MoE model on positive text–image pairs using contrastive learning. This stage aims to build a robust general alignment between textual and visual representations across the diverse domains.

Following this, we perform an intermediate step where we test the model on the training set itself. This process allows us to efficiently mine hard negative samples (i.e., images that are semantically incorrect but have high similarity scores) for each query.

In Stage 2 (Hard-Negative Refinement), we retrain the model, this time providing it with triplets of (query text, positive image, hard negative image). This stage sharpens the model's discriminative ability, forcing it to learn the subtle differences between correct and highly similar incorrect images. This progressive strategy significantly improves model robustness under heterogeneous domains without increasing the total parameter count.

Inference Process During inference, for a given text query q and a candidate image I, the final similarity score is computed as a dynamically weighted sum of the individual expert scores. The entire process is formalized in Equation 1.

$$S_{\text{final}}(q, I) = \sum_{k \in \{\text{sat, drone, ground}\}} g_k(q) \cdot S_k(q, I) \quad (1)$$

All candidate images in the gallery are ranked based on this final score S_{final} to produce the retrieval results.

4. Experiments

This section presents a series of experiments designed to validate the efficacy of our proposed PE-MoE framework. We detail our experimental setup, present our main results in the competition, and conduct in-depth ablation studies to analyze the contribution of each component.

4.1. Experimental Setup

Dataset We use the official data provided by the *RoboSense Challenge 2025* [69] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [70, 71] at ICRA 2023 and the *RoboDrive Challenge 2024* [72, 73] 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 [11, 12, 74–77]. Specifically, this task is built upon the **GeoText-1652** dataset [11] in **Track 4**, which benchmarks cross-modal image-text retrieval for language-guided drone navigation across drastically different viewpoints and real-world sensing conditions.

Evaluation Metrics We adopted the official evaluation metrics for the challenge, which are Recall at K (R@K) for K=1, 5, and 10. R@K measures the percentage of queries for which the correct gallery image is retrieved within the top K results.

Table 1. Ablation analysis of the components in our proposed framework.

#	Model Configuration	R@1	R@5	R@10	Score
1	Baseline: Unified Model w/o Preprocessing	21.32	35.90	42.01	31.67
2	+ Textual Domain Alignment	27.87	45.13	53.22	40.55
3	+ Static Ensemble of Expert Heads	34.42	49.77	58.23	46.33
4	**Full Model: PE-MoE w/ Dynamic Gating**	**38.31**	**53.70**	**61.32**	**49.82**

Table 2. Performance comparison on the University-1652 test set leaderboard.

Method	R@1	R@5	R@10	Score
Official Baseline	25.44	40.61	49.10	39.27
2nd Place	28.34	54.08	66.11	47.23
3rd Place	31.33	49.09	57.15	44.24
Our PE-MoE	**38.31**	**53.70**	**61.32**	**49.82**

4.2. Implementation Details

Our framework was implemented in PyTorch. The shared backbones were initialized from the pre-trained weights of `bge-m3-base` and `eva-clip-large`. Each expert head consisted of the final two trainable transformer layers of text encoder, the final four trainable transformer layers of image encoder and a linear projection layer to map visual features to a 1024-dimensional space. The gating network was a 2-layer MLP with a 512-dimensional hidden layer. We used the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of 1×10^{-4}. All images were resized to 384×384 pixels. The models were trained on eight NVIDIA A100 (80GB) GPUs with a batch size of 128.

4.3. Main Results

Our proposed PE-MoE framework achieved state-of-the-art performance on the official test set, securing first place on the final leaderboard. Table 2 presents a comparison of our results against the official baseline and other top-performing teams. The results clearly demonstrate the superiority of our approach across all key metrics.

4.4. Ablation Study

To rigorously evaluate the contribution of each component in our framework, we conducted a comprehensive ablation study. We started with a basic unified model and progressively added our proposed techniques. The results are summarized in Table 1.

Analysis The results from our ablation study lead to several key insights. First, comparing model #2 to #1, the introduction of our textual domain alignment strategy yields a significant improvement in R@1, confirming its crucial role in mitigating the text domain gap. Second, the transition from model #2 to #3, which replaces the unified model with specialized expert heads (fused with static weights), results in another substantial performance leap. This validates our core "divide and conquer" hypothesis. Finally, comparing our full model (#4) to the static ensemble (#3), the dynamic gating network provides a further discernible boost in accuracy. This demonstrates that an intelligent, query-aware routing mechanism is superior to a fixed-weight fusion, allowing the system to adaptively leverage the best expert for each specific query. Together, these components synergistically contribute to the overall state-of-the-art performance of our final model.

5. Conclusion

In this work, we presented a winning solution to RoboSense 2025 Track 4: Cross-Modal Drone Navigation. To address the challenges of severe platform heterogeneity and textual domain gaps, we proposed a Parameter-Efficient Mixture-of-Experts (PE-MoE) framework combined with a domain-aligned preprocessing pipeline. Specifically, our approach partitions data by platform, augments scarce satellite imagery while sanitizing captions, and aligns the training text distributions via sentence-level splitting. Built upon frozen pre-trained encoders (BGE-M3 and EVA-CLIP), lightweight expert heads specialize in distinct platforms, and a dynamic gating network adaptively routes queries for optimal retrieval. Extensive experiments on the official benchmark demonstrated that our framework achieves state-of-the-art performance and ranked first on the leaderboard, validating its robustness and effectiveness in heterogeneous cross-modal geo-localization. Looking forward, future research may focus on developing end-to-end trainable MoE frameworks, exploring dynamic routing strategies beyond simple softmax gating, and integrating multi-scale and temporal cues for enhanced UAV navigation in complex, real-world environments.

References

[1] Xin Zhou, Xuerong Yang, and Yanchun Zhang. Cdm-net: A framework for cross-view geo-localization with multimodal data. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[2] Ze Song, Xudong Kang, Xiaohui Wei, Shutao Li, and Haibo Liu. Unified and

real-time image geo-localization via fine-grained overlap estimation. *IEEE Transactions on Image Processing*, 2024.

[3] Junyan Ye, Honglin Lin, Leyan Ou, Dairong Chen, Zihao Wang, Qi Zhu, Conghui He, and Weijia Li. Where am i? cross-view geo-localization with natural language descriptions. *arXiv preprint arXiv:2412.17007*, 2024.

[4] Zeying Gong et al. Stairway to success: An online floor-aware zero-shot object-goal navigation framework via LLM-driven coarse-to-fine exploration. *arXiv preprint arXiv:2505.23019*, 2025.

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

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

[7] Xupei Zhang, Hanlin Qin, Lin Ma, Yue Yu, Yang Ma, and Yanhao Hu. Deep feature matching of different-modal images for visual geo-localization of uavs. *IEEE Transactions on Aerospace and Electronic Systems*, 2024.

[8] Jiahao Wen, Hang Yu, and Zhedong Zheng. Weatherprompt: Multi-modality representation learning for all-weather drone visual geo-localization. *arXiv preprint arXiv:2508.09560*, 2025.

[9] 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.

[10] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1472–1484, 2025.

[11] 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.

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

[13] GuangHao Meng, Sunan He, Jinpeng Wang, Tao Dai, Letian Zhang, Jieming Zhu, Qing Li, Gang Wang, Rui Zhang, and Yong Jiang. Evdclip: Improving vision-language retrieval with entity visual descriptions from large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6126–6134, 2025.

[14] Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-clip: Multimodal continual learning for vision-language model. In *The Thirteenth International Conference on Learning Representations*, 2025.

[15] Yiguo He, Junjie Zhu, Yiying Li, Qiangjuan Huang, Zhiyuan Wang, and Ke Yang. Rethinking remote sensing clip: Leveraging multimodal large language models for high-quality vision-language dataset. In *International Conference on Neural Information Processing*, pages 417–431. Springer, 2024.

[16] Ali Asgarov and Samir Rustamov. Lowclip: Adapting the clip model architecture for low-resource languages in multimodal image retrieval task. *arXiv preprint arXiv:2408.13909*, 2024.

[17] Mohammed Elhenawy, Huthaifa I Ashqar, Andry Rakotonirainy, Taqwa I Alhadidi, Ahmed Jaber, and Mohammad Abu Tami. Vision-language models for autonomous driving: Clip-based dynamic scene understanding. *Electronics*, 14(7):1282, 2025.

[18] 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.

[19] Runnan Chen et al. Towards label-free scene understanding by vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pages 75896–75910, 2023.

[20] Jingyi Xu et al. Visual foundation models boost cross-modal unsupervised domain adaptation for 3d semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 26(11):20287–20301, 2025.

[21] Xidong Peng, Runnan Chen, Feng Qiao, et al. Learning to adapt SAM for segmenting cross-domain point clouds. In *European Conference on Computer Vision*, pages 54–71. Springer, 2024.

[22] Youquan Liu et al. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pages 37193–37229, 2023.

[23] Shuo Xing, Yuping Wang, Peiran Li, Ruizheng Bai, Yueqi Wang, Chan-wei Hu, Chengxuan Qian, Huaxiu Yao, and Zhengzhong Tu. Re-align: Aligning vision language models via retrieval-augmented direct preference optimization. *arXiv preprint arXiv:2502.13146*, 2025.

[24] Leqi Shen, Guoqiang Gong, Tianxiang Hao, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, Jungong Han, and Guiguang Ding. Discovla: Discrepancy reduction in vision, language, and alignment for parameter-efficient video-text retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19702–19712, 2025.

[25] Jinhao Li, Haopeng Li, Sarah Erfani, Lei Feng, James Bailey, and Feng Liu. Visual-text cross alignment: Refining the similarity score in vision-language models. *arXiv preprint arXiv:2406.02915*, 2024.

[26] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[27] Xiang Xu, Lingdong Kong, Hui Shuai, Liang Pan, Ziwei Liu, and Qingshan Liu. LiMoE: Mixture of LiDAR representation learners from automotive scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27368–27379, 2025.

[28] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[29] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[30] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024.

[31] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13019–13029, 2024.

[32] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024.

[33] Kaicheng Yang, Tiancheng Gu, Xiang An, Haiqiang Jiang, Xiangzi Dai, Ziyong Feng, Weidong Cai, and Jiankang Deng. Clip-cid: Efficient clip distillation via cluster-instance discrimination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21974–21982, 2025.

[34] Chenyang Yu, Xuehu Liu, Yingquan Wang, Pingping Zhang, and Huchuan Lu. Tf-clip: Learning text-free clip for video-based person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 6764–6772, 2024.

[35] Lingdong Kong, Xiang Xu, Jiawei Ren, et al. Multi-modal data-efficient 3D scene understanding for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3748–3765, 2025.

[36] 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.

[37] 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.

[38] Pengfei Wei et al. Unsupervised video domain adaptation for action recognition: A disentanglement perspective. In *Advances in Neural Information Processing Systems*, volume 36, pages 17623–17642, 2023.

[39] Xuzhi Wang, Xinran Wu, Song Wang, et al. Monocular semantic scene completion via masked recurrent networks. In *IEEE/CVF International Conference on Computer Vision*, pages 24811–24822, 2025.

[40] 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.

[41] Hailang Huang, Zhijie Nie, Ziqiao Wang, and Ziyu Shang. Cross-modal and uni-modal soft-label alignment for image-text retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18298–18306, 2024.

[42] Yongquan Wan, Wenhai Wang, Guobing Zou, and Bofeng Zhang. Cross-modal feature alignment and fusion for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8384–8388, 2024.

[43] Gang Hu, Zaidao Wen, Yafei Lv, Jianting Zhang, and Qian Wu. Global–local information soft-alignment for cross-modal remote-sensing image–text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[44] Tiantian Gong, Junsheng Wang, and Liyan Zhang. Cross-modal semantic aligning and neighbor-aware completing for robust text–image person retrieval. *Information Fusion*, 112:102544, 2024.

[45] Zhe Li, Lei Zhang, Kun Zhang, Yongdong Zhang, and Zhendong Mao. Improving image-text matching with bidirectional consistency of cross-modal alignment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6590–6607, 2024.

[46] Rock Yuren Pang, Sebastin Santy, René Just, and Katharina Reinecke. Blip: facilitating the exploration of undesirable consequences of digital technologies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.

[47] Muhe Ding, Yang Ma, Pengda Qin, Jianlong Wu, Yuhong Li, and Liqiang Nie. Ra-blip: Multimodal adaptive retrieval-augmented bootstrapping language-image pre-training. *IEEE Transactions on Multimedia*, 2025.

[48] Eren Duman, Oguzhan Serttas, Enes Ozelbas, and Ali Can Karaca. Blip-cc: Adapting the blip for change captioning task in remote sensing. In *2025 33rd Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2025.

[49] Yining Huang, Zuobai Zhang, Jian Tang, Debora Susan Marks, and Pascal Notin. Augmenting evolutionary models with structure-based retrieval. In *ICML'24 Workshop ML for Life and Material Science: From Theory to Industry Applications*, 2024.

[50] Minjun Cho, Sungwoo Kim, Dooho Choi, and Yunsick Sung. Enhanced blip-2 optimization using lora for generating dashcam captions. *Applied Sciences*, 15(7):3712, 2025.

[51] Wei Chen, Changyong Shi, Chuanxiang Ma, Wenhao Li, and Shulei Dong. Depthblip-2: Leveraging language to guide blip-2 in understanding depth information. In *Proceedings of the Asian Conference on Computer Vision*, pages 2939–2953, 2024.

[52] Yunzhe Xiao, Yong Dou, and Shaowu Yang. Pointblip: Zero-training point cloud classification network based on blip-2 model. 2024.

[53] Matheus Fernandes de Sousa. Aplicação do modelo de linguagem blip-2 na geração automática de descrições em vídeos esportivos. 2024.

[54] Ze Gao, Jing Guo, Liming Chen, Kai Wang, Yang Chen, Yongzhen Ke, and Shuai Yang. Andr-blip2: Enhanced semantic understanding framework for industrial image anomaly detection and report generation. *Journal of the Franklin Institute*, page 107816, 2025.

[55] Qian Huang, Xiaotong Guo, Yiming Wang, Huashan Sun, and Lijie Yang. A survey of feature matching methods. *IET Image Processing*, 18(6):1385–1410, 2024.

[56] Qian Huang, Xiaotong Guo, Yiming Wang, Huashan Sun, and Lijie Yang. A survey of feature matching methods. *IET Image Processing*, 18(6):1385–1410, 2024.

[57] Zeyu Ma, Yuqi Li, Yizhi Luo, Xiao Luo, Jinxing Li, Chong Chen, Xian-Sheng Hua, and Guangming Lu. Discrepancy and structure-based contrast for test-time adaptive retrieval. *IEEE Transactions on Multimedia*, 26:8665–8677, 2024.

[58] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021.

[59] Francesco Pro, Nikolaos Dionelis, Luca Maiano, Bertrand Le Saux, and Irene Amerini. A semantic segmentation-guided approach for ground-to-aerial image matching. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 2630–2635. IEEE, 2024.

[60] Emanuele Mule, Matteo Pannacci, Ali Ghasemi Goudarzi, Francesco Pro, Lorenzo Papa, Luca Maiano, and Irene Amerini. Enhancing ground-to-aerial image matching for visual misinformation detection using semantic segmentation. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 795–803, 2025.

[61] Chaoran Li, Chao Yan, Xiaojia Xiang, Jun Lai, Han Zhou, and Dengqing Tang. Ample: Automatic progressive learning for orientation unknown ground-to-aerial geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[62] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.

[63] Shiyi Cao, Shu Liu, Tyler Griggs, Peter Schafhalter, Xiaoxuan Liu, Ying Sheng, Joseph E Gonzalez, Matei Zaharia, and Ion Stoica. Moe-lightning: High-throughput moe inference on memory-constrained gpus. In *ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 715–730, 2025.

[64] Xiang Xu et al. FRNet: Frustum-range networks for scalable LiDAR segmentation. *IEEE Transactions on Image Processing*, 34:2173–2186, 2025.

[65] Xiang Xu et al. 4D contrastive superflows are dense 3D representation learners. In *European Conference on Computer Vision*, pages 58–80, 2024.

[66] Junyi Chen, Longteng Guo, Jia Sun, Shuai Shao, Zehuan Yuan, Liang Lin, and Dongyu Zhang. EVE: Efficient vision-language pre-training with masked prediction and modality-aware MoE. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 1110–1119, 2024.

[67] Xiaoda Yang, JunYu Lu, Hongshun Qiu, Sijing Li, Hao Li, Shengpeng Ji, Xudong Tang, Jiayang Xu, Jiaqi Duan, Ziyue Jiang, et al. Astrea: A moe-based visual understanding model with progressive alignment. *arXiv preprint arXiv:2503.09445*, 2025.

[68] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[69] 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, Yuchen Yang, 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 Su, 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.

[70] 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.

[71] 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*, volume 36, pages 21298–21342, 2023.

[72] 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, et al. The RoboDrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.

[73] 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.

[74] 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.

[75] 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.

[76] 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*, volume 37, pages 34980–35017, 2024.

[77] 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.