Turning Generators into Retrievers: Unlocking MLLMs for Natural Language-Guided Geo-Localization

Supplementary Material

1. Introduction

This supplementary document provides comprehensive details supporting the method and findings presented in the main paper. We organize our supplementary materials as following:

- Section 2 Implementation Details. We present the specific parameters and environmental configurations used for our experiments, offering detailed insight into the reproducibility of our work.
- Section 3 Generalization and Robustness. We evaluate our model's zero-shot cross-city performance and its robustness to data scarcity, demonstrating superior spatial reasoning with minimal trainable parameters.
- Section 4 Necessity to Finetune MLLMs. We demonstrate the importance of task-specific finetuning for adapting MLLMs to the NGCG task.
- Section 5 Qualitative Results. We present qualitative results from the CVG-Text benchmark, including both successful and unsuccessful text-to-satellite retrieval examples for Brisbane and Tokyo.
- Section 6 Societal Impact. We discuss the broader implications of the proposed model, including ethical considerations, potential biases, and strategies for responsible deployment.

2. Implementation Details

We choose InternVL3.5-1B [4], SmolVLM-500M [2], and SmolVLM-256M [2] as our MLLM backbone to compare performances across different model scales. The comparative results are reported in the main manuscript. The model is trained by the Adam optimizer [1] with a learning rate of $3e^{-5}$ and the cosine learning rate scheduler. The input image resolution is resized to the MLLM's pre-defined image input size, and the maximum textual context length is fixed at 300 tokens. We set the temperature coefficient τ in the infoNCE [3] loss to 0.03. The batch size is set to 12. The epoch is set to 20 for GeoText-1652 and 100 for CVGText, aligning with approximately 200,000 training steps. We utilize LoRA fine-tuning with $r = 16$ and set the $\frac{\alpha}{r}$ ratio to 8. The training was conducted on a single NVIDIA H100 GPU.

3. Generalization and Robustness to Data Scarcity

To verify generalizable spatial reasoning, we evaluate zero-shot cross-city performance on CVGText OSM (Table 1). Despite updating only 9M parameters, our method outperforms the fully fine-tuned CrossText2Loc (428M) in 4/6 scenarios. Furthermore, we clarify that the Brisbane OSM performance drop (*Table 3 of the main paper*) stems from inherent local data noise, not overfitting. Notably, on the more complex Brisbane satellite task, Ours-I achieves 45.25% R@1 vs. 43.58% for CrossText2Loc, demonstrating robust reasoning in data-scarce regions.

Table 1. Zero-shot cross-city generalization on OSM data. Source cities are indicated in bold.

	NewYork		**Tokyo**		**Brisbane**	
Method	Brisbane	Tokyo	NewYork	Brisbane	NewYork	Tokyo
CrossText2Loc	25.67	18.58	**41.08**	**26.00**	34.92	14.67
Ours-I	**26.25**	**22.00**	39.00	25.16	**37.33**	**19.00**

4. Necessity to Finetune MLLMs

To validate the necessity of task-specific adaptation for MLLMs in Natural Language-Guided Cross-view Geolocalization (NGCG), we compare the performance with finetuning and without finetuning on InternVL3.5-1B [4] backbone. We first report results on the complete New York text-to-satellite retrieval task, which contains 1,200 samples. The results in Table 2 confirm the hypothesis presented in the introduction that **the representations learned by MLLMs that are optimized for next-token prediction lack the discriminative properties required for retrieval.**. Without finetuning, which uses the pretrained InternVL3.5 directly for retrieval, achieves an R@1 score of only 0.92% and an L@50 accuracy of 0.92%. This performance confirms that the MLLM's latent space, despite general cross-modal alignment, is not naturally structured for geographically grounded cross-modal matching. The application of our retrieval-specific fine-tuning strategy yields a dramatic performance increase. By applying our finetuning techniques achieve an R@1 of 44.75% and an R@10 of 86.00%. This significant shift demonstrates that our fine-tuning framework reshapes the MLLM's embedding space for precise NGCG retrieval performance.

Table 2. Overall Performance Comparison of with and without fine-tuning MLLMs for NGCG. We report recall rates (%) and localization hit rates (%) based on thresholded retrieval. 'FT' is short for fine-tuning. [Key: **Best**]

Strategy	R@1	R@5	R@10	L@50	L@100	L@150
Without FT	0.92	5.00	9.50	0.92	1.75	3.75
With FT	**44.75**	**75.92**	**86.00**	**47.50**	**51.33**	**54.25**

Table 3. Comparative Retrieval and Localization Performance by Two Input View Types (Panorama and Single-View). We report recall rates (%) and localization hit rates (%) based on thresholded retrieval. 'FT' is short for fine-tuning. [Key: **Best**]

Strategy	Panorama(1000 samples)						Single-View(200 samples)					
	R@1	R@5	R@10	L@50	L@100	L@150	R@1	R@5	R@10	L@50	L@100	L@150
Without FT	1.10	5.80	11.10	1.10	1.90	4.00	0.00	1.00	1.50	0.00	1.00	2.50
With FT	**49.80**	**83.40**	**92.90**	**52.80**	**56.90**	**59.40**	**19.50**	**38.50**	**51.50**	**21.00**	**23.50**	**28.50**

We separate 1000 panorama-view samples and 200 single-view samples from the New York text-to-satellite retrieval task and report their performance in Table 3. Without finetuning, the performance is poor on both views, and especially underperformed for the single-view data. The finetuned model performs optimally on the panorama subset, achieving an R@1 of 49.80% and R@5 of 83.40%. The extended angular context provided by the panorama images allows the model to effectively capture surrounding landmarks, road structures, and environmental cues necessary for highly reliable retrieval. Performance drops notably on the single-view subset, with an R@1 of 19.50%. This outcome demonstrates the model's dependency on wide-angle context, even in the text space. Retrieving a location based on a narrow perspective is more challenging, as the limited field of view often omits crucial disambiguating features mentioned in the query text. **This highlights single-view localization as a key area for future work**.

5. Qualitative Results

In this section, we provide additional successful and failed retrieval results from the CVG-Text benchmark for the Brisbane and Tokyo regions. The retrieval results for New York are presented in the main manuscript.

5.1. Successful Retrievals

We provide additional qualitative results of text-to-satellite image retrieval on the **Brisbane** dataset in Figure 1. The results demonstrate our model's capability to localize images based on a combination of functional, named, and structural features. In the first example, the query demonstrates correct alignment based on functional land-use features and specific, named architectural landmarks. The model uses the high contrast between residential/green space and the modern building to achieve high localization accuracy. The second query shows a crucial challenge related to spatial variance. While the ground truth is ranked Top 1, the presence of the incorrect Top 2 image, which is slightly shifted from the Top 1 location, indicates that the model successfully identified the core visual features, such as intersection geometry and red-brick structure. The third query, containing several specific named entities ("post office", "Drakos Solicitors", "Kurilpa Senior Citizens Centre") confirms the model's capacity to localize based on a dense collection of commercial and civic buildings. **This highlights the model's ability to aggregate multiple descriptive cues to find the correct location, despite the potential visual ambiguity of individual residential structures in the proximity. More importantly, this outcome confirms that our MLLM fine-tuning approach is indeed the correct path, as it effectively leverages the MLLM's inherent cross-modal alignment ability to grasp the nuanced semantics required for the NGCG task.**

Additional visualization on the **Tokyo** subset is illustrated in Figure 2, which shows that our model effectively correlates fine-grained urban features described in text with their corresponding satellite views. For the first example, the model successfully identifies the high-level geometric layout: a broad multi-lane road immediately bordered by a large, distinct open space. The second example, featuring complex road markings and clear pedestrian crossings, shows the model can accurately map low-level road topology and traffic flow patterns onto the satellite image. The third query sug-

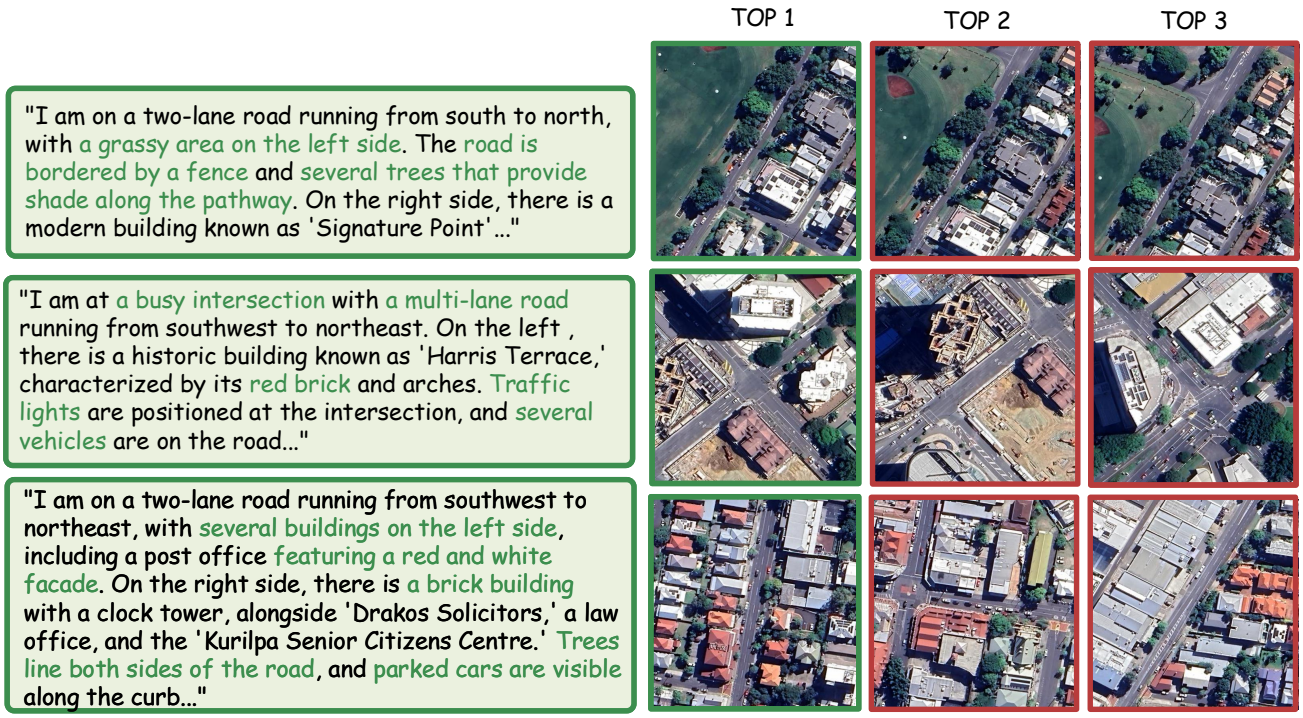

Figure 1. Visualization of Text-to-Satellite Image Retrieval Results on CVG-Text Brisbane. Three text-satellite pairs are shown. From left to right is the query text, the retrieved satellite images, and the ground truth satellite images. Ground truth satellite images are outlined with a green border, while incorrect matches are outlined with a red border.

gests the model can align text to the unique, distinctive architecture of large urban structures visible from the satellite view. Crucially, even when a specific feature like the "curved design" might appear in a visually similar, incorrect Top 2 candidate, the model maintains high retrieval accuracy. **This demonstrates the framework's ability to resolve semantic ambiguity by successfully integrating and weighting all complementary cues**, such as the specific orientation of the curved building relative to the road and the surrounding landscaping, to select the correct Ground Truth image.

5.2. Failure Cases Analysis

We first report **Brisbane** failure retrieval results in Figure 3. A crucial finding from our retrieval analysis is that the observed TOP 1 failures in the first and second rows stem from the limitations imposed by an overly strict positive sample definition within the evaluation protocol. It only counts the exact file of the correct satellite image as a success. Because the image files are named by their latitude and longitude, the model's TOP 1 result was marked as a failure, even though it showed the same geographical location, just from a slightly different viewing angle or as an adjacent map tile. Consequently, the correct location was retrieved at TOP 1, but the result was penalized due to non-matching reference coordinates. This indicates that our model exhibits a stronger fine-grained localization capability than suggested by the R@1 metric, as it successfully guided the query embedding to the correct region.

The performance on **Tokyo** illustrated in Figure 4 shows why we got lower performance at Tokyo when comparing it with other cities. This might be because the standard tokenizer fails to maintain the semantic integrity of Japanese like 東京ビルヂング (Tokyo Building), which causes the output embeddings for the query to be semantically weak. Similarly, in the second query, our model successfully performs coarse-grained scene recognition, identifying general features such as road landmarks and the presence of a building structure. However, it exhibits a critical deficiency in fine-grained visual grounding, specifically lacking the capability for precise verification required to confirm the identity of the building. The third query failed due to a lack of contextual data. Without OpenStreetMap details, the model could only form a generic query based on local visual cues like "paved ground" and "narrow alleyway," which was insufficient to disambiguate be-

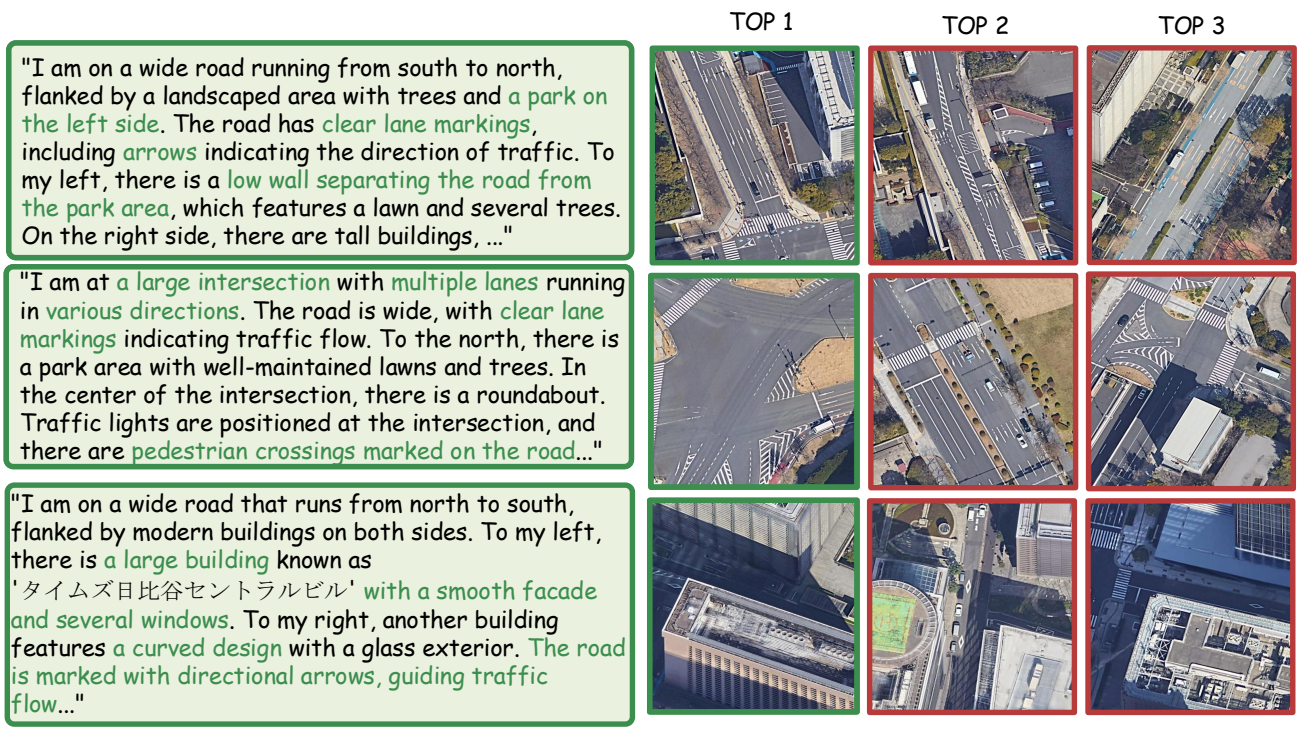

Figure 2. Visualization of Text-to-Satellite Image Retrieval Results on CVG-Text Tokyo. Three text-satellite pairs are shown. Left is the query text. On the right is the the retrieved satellite images. Ground Truth satellite images are outlined with a green border, while incorrect matches are outlined with a red border.

tween visually similar locations.

In summary, **the observed failure cases indicate that our model is sometimes constrained by the original MLLMs.** For example, the standard tokenization process can occasionally misunderstand the full semantic information of complex geographic proper nouns, leading to a slight drop in precise embedding for the search query. Likewise, while the model's visual pipeline performs reliably on general scene understanding, it remains limited when confronted with the fine-grained verification tasks required to disambiguate visually similar landmarks. **These challenges, especially in visually ambiguous or semantically dense scenarios, can be effectively alleviated by scaling up to more advanced, heavier MLLMs, such as Qwen, GLM, or LLaVA.** Such state-of-the-art models, which are now readily available, naturally provide stronger semantic preservation and more discriminative visual–textual reasoning due to their larger capacity, richer tokenization strategies, and improved architectural components.

6. Societal Impact

Natural language Guided Cross-view Geo-localization (NGCG) aims to map a free-form natural language description to its corresponding satellite-view location. This ability is especially valuable in environments where GPS signals are weak, such as urban canyons. For emergency responders, NGCG can assist dispatchers in interpreting descriptive information from callers and narrowing down likely locations, potentially reducing response latency when precise GPS coordinates are not readily available. NGCG also offers broader benefits across civilian applications, such as urban planning and commercial development.

For consumer platforms such as Yelp or Google Maps, NGCG enables users to search using detailed visual or contextual attributes described in language. For example, a user might specify "a restaurant with a rustic interior, exposed brick, and high ceilings." and NGCG can match this description to candidate locations by associating the text with satellite imagery or linked interior/exterior photos. This allows geospatial search to go beyond coarse category filters toward more personalized, visually grounded discovery.

Similarly, delivery platforms can leverage NGCG to resolve last-mile and short-distance delivery issues that standard mapping addresses cannot handle. If a standard address is vague, a delivery driver's specific natural language observation ("The package is next to

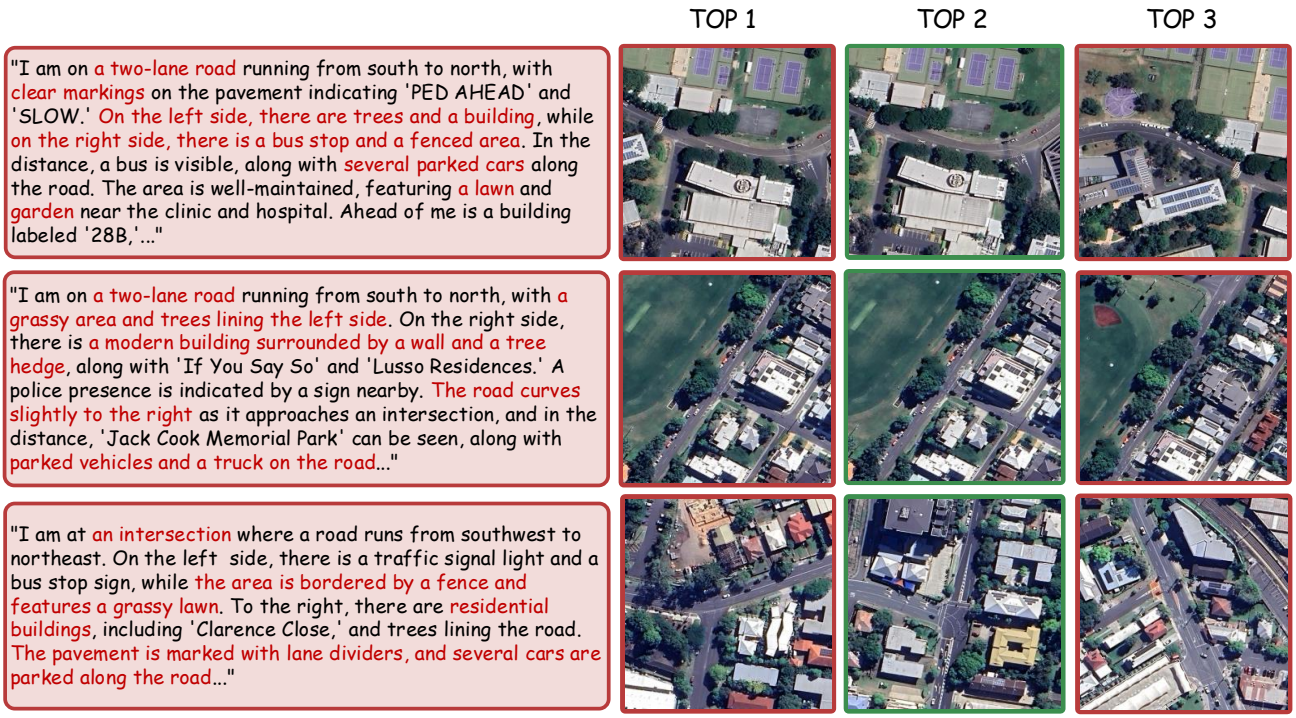

Figure 3. Visualization of Text-to-Satellite Image Retrieval Results on the CVG-Text Brisbane subset. Three query pairs are shown where the correct match was retrieved at TOP 2, demonstrating a retrieval failure at TOP 1. The left column contains the natural language query text. The right columns display the retrieved satellite images, where the Ground Truth image is outlined in green and incorrect matches are outlined in red.

the green mailbox under the porch light") can be reasoned against the recorded street view to confirm the placement. This reduces failed deliveries, minimizes the need for follow-up calls, and consequently improves overall supply chain efficiency.

The real-time reasoning capability inherent in the MLLM framework significantly strengthens urban resilience. The system can reason about environmental and infrastructure changes in real-time. For instance, an MLLM could analyze a satellite image and a descriptive report ("The construction has blocked off two lanes on Main Street since Monday") to dynamically update route planning. This capability helps citizens avoid congestion, improves the overall flow of city traffic, and contributes to the real-time operational efficiency necessary for urban resilience.

In summary, while NGCG has many potential benefits, it also needs to be utilized carefully. Systems should protect people's privacy information, prevent misuse, and avoid situations where an incorrect location might cause problems, especially in emergencies. With these safeguards, **NGCG can make people's lives easier to interact with maps and location services, improving safety, local business search, delivery services, and everyday navigation**. Moreover, it is also necessary to include more diverse geographical areas into account, for instance, suburban areas, rural areas, and mountainous areas, extending the coverage and usability of NGCG models.

References

[1] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[2] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 1

[3] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1

[4] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1

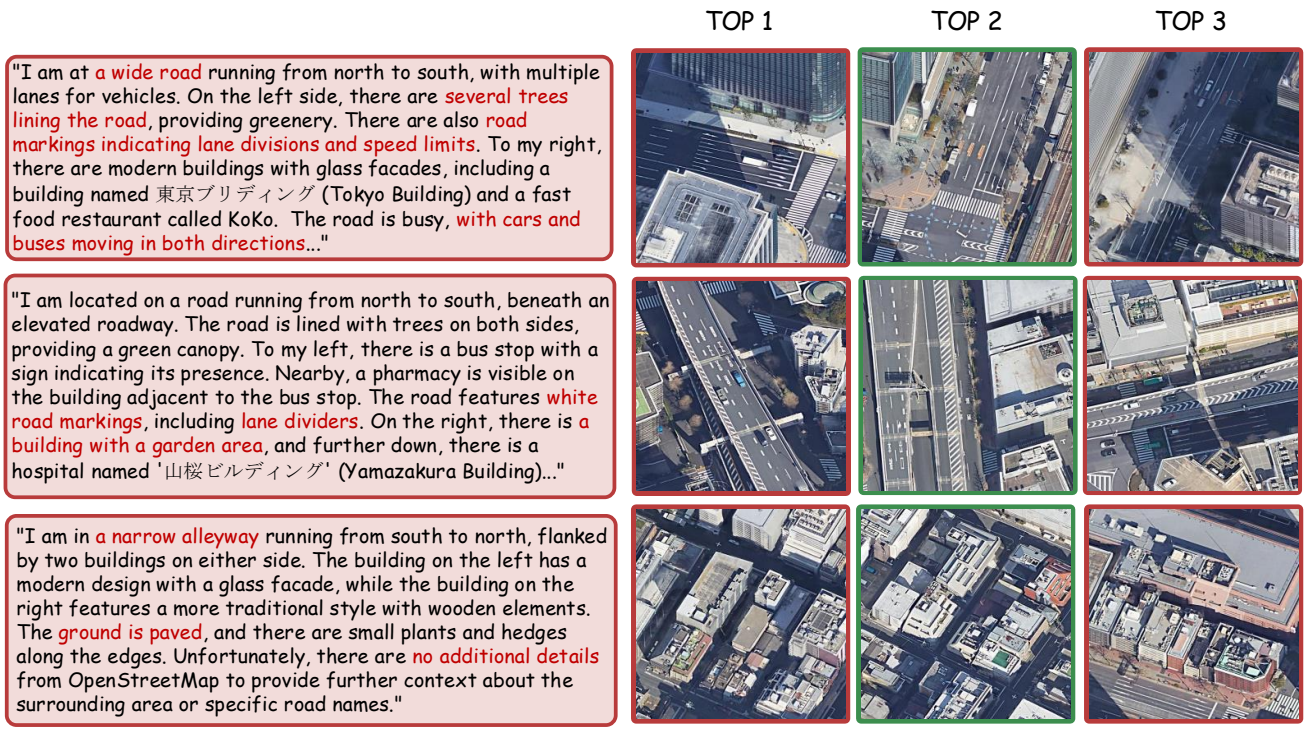

Figure 4. Visualization of Text-to-Satellite Image Retrieval Results on the CVG-Text Tokyo subset. Three query pairs are shown where the correct match was retrieved at TOP 2, demonstrating a retrieval failure at TOP 1. The left column contains the natural language query text. The right columns display the retrieved satellite images, where the Ground Truth image is outlined in green and incorrect matches are outlined in red.