

Where am I? Cross-View Geo-localization with Natural Language Descriptions

Junyan Ye^{1,2*}, Honglin Lin^{2*}, Leyan Ou¹,
Dairong Chen^{4,1}, Zihao Wang¹, Qi Zhu¹, Conghui He^{2,3}, Weijia Li^{1†}

¹Sun Yat-Sen University, ²Shanghai AI Laboratory, ³Sensetime Research, ⁴Wuhan University

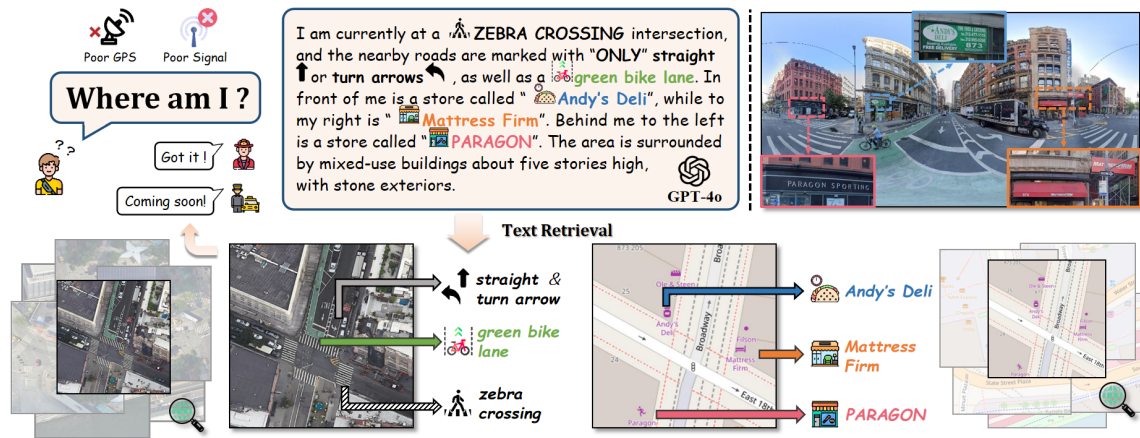


Figure 1. **Towards Text-guided Geo-localization.** In scenarios where GPS signals are interfered with, users must describe their surroundings using natural language, providing various location cues to determine their position (*Up*). To address this, we introduce a text-based cross-view geo-localization task, which retrieves satellite imagery or OSM data only based on text queries for position localization (*Down*).

Abstract

Cross-view geo-localization identifies the locations of street-view images by matching them with geo-tagged satellite images or OSM. However, most existing studies focus on image-to-image retrieval, with fewer addressing text-guided retrieval, a task vital for applications like pedestrian navigation and emergency response. In this work, we introduce a novel task for cross-view geo-localization with natural language descriptions, which aims to retrieve corresponding satellite images or OSM database based on scene text descriptions. To support this task, we construct the CVG-Text dataset by collecting cross-view data from multiple cities and employing a scene text generation approach that leverages the annotation capabilities of Large Multimodal Models to produce high-quality scene text descriptions with localization details. Additionally, we propose a novel text-based retrieval localization method, CrossText2Loc, which demonstrates excellent long-text retrieval capabilities. In terms of explainability, it not only provides similarity scores but also offers retrieval reasons. More can be found at <https://github.com/yejy53/CVG-Text>.

1. Introduction

Accurate positioning of ground-level images is crucial for various applications, including pedestrian navigation [35], mobile robot localization [32], and noisy GPS signals correction in crowded urban areas [16, 37]. Traditional localization methods typically rely on 3D point cloud positioning [17, 26] or cross-view retrieval using GPS-tagged satellite images [10, 44, 45]. However, most research focuses on matching image-to-3D data or image-to-image data. Recent studies have begun to explore a novel localization approach based on natural language text, which holds significant value for many practical applications [35, 39, 40]. As shown in Figure 1, taxi drivers rely on verbal instructions from passengers to determine their location [21], or pedestrians describe their position during emergency calls [7].

Recent natural language localization methods, such as Text2Pose[21] and Text2Loc [40], have been limited to using natural language to identify individual locations within point clouds. Constructing 3D maps using LiDAR or photogrammetry is expensive on a global scale [1, 13, 27], and the storage costs for 3D maps are also high, often requiring costly cloud infrastructure, which hinders localization on mobile devices. Notably, the cross-view retrieval geo-

localization paradigm [41, 45, 56] that utilizes OSM¹ map data or satellite imagery, although oriented towards coarse-grained localization, can still meet the needs of most tasks and has clear advantages over 3D data in terms of coverage and storage costs. Therefore, this paper introduces a novel cross-view geo-localization task, i.e., exploring the use of natural language descriptions to retrieve corresponding OSM or satellite images.

To tackle this challenging task, a dataset is needed that (i) contains foundational cross-view data with street-view, OSM, and satellite images, and (ii) includes text data capable of simulating human users in describing street-view scenes, while providing high-quality scene localization cues. With the development of large multimodal models (LMMs), annotating text using LMMs seems to be an effective solution [8, 14, 46]. However, LMMs may suffer from vague descriptions or hallucination phenomena [15, 31]. To address these issues, we propose the Cross-View Geo-localization dataset, CVG-Text. We first collected street-view data from over 30,000 locations in three cities, New York, Brisbane, and Tokyo. Then, based on the geographical coordinates, we obtained corresponding paired data of OSM and satellite images. Subsequently, we developed a progressive text description framework that leverages LMM, GPT-4o [28] as the core for generation, combining Optical Character Recognition (OCR) and Open-World Segmentation [29, 52] techniques to generate high-quality scene description text from street-view images while reducing vague descriptions.

Although the textual data constructed above can provide user-like street scene descriptions, it still has still have a significant domain gap compared to satellite images or OSM data. Moreover, in order to fully capture the scene’s detailed information, the generated text descriptions are generally long, often exceeding the text encoding limits of image-text retrieval methods. To address this issue, we propose a novel Cross-view Text-based Localization method, *CrossText2Loc*. This method includes a length-extended text encoding module, Extended Embedding, which fully leverages the long and complex text descriptions in the dataset. Through contrastive learning strategies, it effectively learns cross-domain matching information. It also features an Explainable Retrieval Module (ERM), which provides natural language explanations alongside the retrieval results. This overcomes the limitations of traditional cross-view retrieval methods that only provide similarity scores, lacking interpretability and making it difficult to make confident decisions. We evaluate the performance of mainstream text-image retrieval methods and our method on this novel task, and the experimental results demonstrate that our *CrossText2Loc* has significant advantages in recall metrics and interpretability. Our main contributions are as follows:

- We introduce and formalize the Cross-View Geo-

localization task based on natural language descriptions, utilizing scene text descriptions to retrieve corresponding OSM or satellite images for geographical localization.

- We propose *CVG-Text*, a dataset with well-aligned street-views, satellite images, OSM, and text descriptions across three cities and over 30,000 coordinates. Additionally a progressive scene text generation framework based on LMM is presented, which reduces vague descriptions and generates high-quality scene text.
- We introduce *CrossText2Loc*, a novel text-based localization method that excels in handling long texts and interpretability. It achieves an improvement of over 10% in Top-1 recall compared to existing methods, while offering retrieval reason and confidence beyond similarity scores.

2. Related Work

2.1. Cross-view Geo-localization

Cross-view geo-localization identifies the geographic locations of street-view images by matching them with geographic reference satellite databases or OSM databases for coarse localization [23, 34, 41, 45]. For example, works like Sample4G [10] employ a contrastive learning framework to match features of street-view images with satellite image features, achieving high-accuracy satellite data retrieval. However, current cross-view retrieval and localization tasks primarily focus on image-to-image, with limited consideration for text, which presents certain shortcomings in practical applications. Additionally, existing cross-view retrieval methods mainly provide similarity score, with little research dedicated to confident and interpretable retrieval localization. Our proposed text retrieval localization task is based on the retrieval and localization of natural language, addressing the gap in existing research regarding text-guided scene localization applications. This approach enables more transparent and interpretable retrieval localization through the use of natural language.

2.2. Visual Language Navigation and Localization

Visual Language Navigation (VLN) requires agents to navigate specific environments based on natural language instructions [2, 6]. Previous tasks have primarily focused on decision-making based on images and natural language. Recent works have begun to shift from visual language navigation to direct visual language localization tasks. For instance, Loc4Plan[35] and AnyLoc[18] emphasize the necessity of visual spatial localization prior to navigation. Text2Pose [21] and Text2Loc [40] explore the use of natural language to identify individual positions in outdoor point cloud maps. However, point cloud retrieval tasks tend to incur high storage and computational costs when handling large-scale areas.

¹<https://www.openstreetmap.org/>

2.3. Data Synthesis via LMMs

The rapid development of multimodal large models (LMMs) has demonstrated their outstanding ability to generate high-quality natural language descriptions [3, 28, 42, 47]. Many studies have leveraged LMMs for automated data annotation [14], such as LatteCLIP [4], which synthesizes text using LMMs for unsupervised CLIP fine-tuning. However, text retrieval localization tasks impose higher demands on the annotation capabilities of LMMs, requiring them to accurately identify key localization details in street-view images, such as store signs and other critical information, while minimizing interference from hallucination phenomena [31]. In our fine-grained text synthesis, we incorporate street-view image, OCR, and open-world segmentation to enhance GPT’s capture capability and reduce hallucination.

2.4. Multi-modality Alignment

In this work, our task involves retrieving satellite images or OSM images based on scene text descriptions synthesized by LMMs, which can be seen as a subtask of text-to-image retrieval [9, 11, 43, 48, 54, 55]. The CLIP [30] model introduced a contrastive learning approach between image-text pairs, establishing a new paradigm for recent text retrieval tasks. However, such models often have fixed maximum sequence length limitations, typically defaulting to 77 tokens, making it challenging to handle complex long-text scene descriptions. This can lead to the loss of fine-grained textual information, negatively impacting model performance. Inspired by works like Long-CLIP [50], our retrieval method employs stretching to extend the model’s acceptable text length, enabling the retrieval of long text descriptions for environmental contexts.

3. CVG-Text Dataset

3.1. Overview

We introduce CVG-Text, a multimodal cross-view retrieval localization dataset designed to evaluate text-based scene localization tasks. CVG-Text covers three cities: New York, Brisbane, and Tokyo, encompassing over 30,000 scene data points. The data from New York and Tokyo is more oriented toward urban environments, while the Brisbane data leans towards suburban scenes. Each individual point includes corresponding street-view images, OSM data, satellite images, and associated scene text descriptions. The dataset is randomly split into training and test sets with a ratio of 5:1. More details can be found in the supplementary materials.

Statistical Overview of Text Data. Figure 2 presents the feature statistics of the scene text. The t-SNE dimensionality reduction visualization indicates a relatively dispersed distribution of the text data, reflecting high diversity. Texts from the same city exhibit clustering, while texts from different cities are clearly distinguishable, highlighting regional varia-

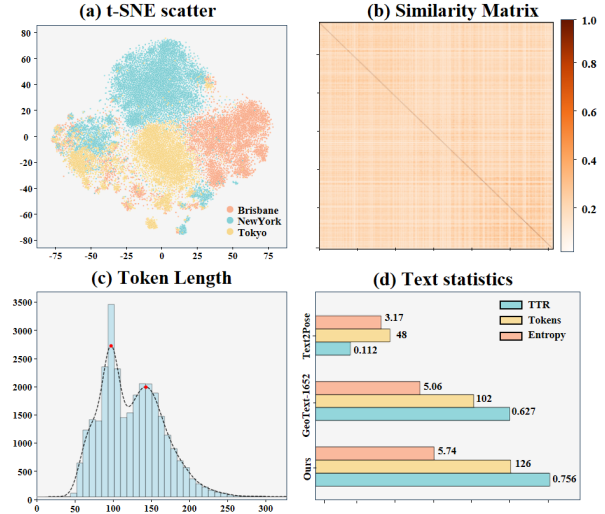


Figure 2. **Textual Feature Statistics Overview.** (a) t-SNE visualization of text data from different cities; (b) text similarity matrix; (c) token length distribution histogram; (d) comparison of text statistics across different datasets.

tions in text features. These differences are likely tied to the unique styles and cultural characteristics of each city. The text similarity matrix reveals low similarity, demonstrating the independence of the texts, which effectively represent each distinct scene and reduce the risk of confusion between different texts. The average text length generated by the multimodal large model GPT-4o exceeds 126 tokens, with two prominent peaks at 100 and 145 tokens, corresponding to single-view and panoramic images, respectively. Compared to datasets such as Text2Pose and GeoText-1652, our data shows superior performance in vocabulary richness, token length, and entropy, reflecting a higher quality of text.

Comparisons with Existing Datasets. Table 1 compares CVG-Text with existing datasets. In contrast to common cross-view retrieval datasets, such as CVUSA [38] and VIGOR [56], CVG-Text includes aligned text modality information, enabling the evaluation of text-based scene localization tasks and interpretability analysis for cross-view retrieval. Furthermore, CVG-Text demonstrates superior data completeness, encompassing panoramic street-views, single-perspective street-views, aerial images, and OSM data.

Compared to GeoText-1652 [9] dataset, which is primarily used for drone navigation, the text descriptions are directly derived from drone images and are used for drone image retrieval. Our task, however, focuses on addressing the needs of pedestrians, tourists, and other users, with text originating from street-view images and used for cross-domain retrieval of satellite or OSM images. There is a significant difference in both task objectives and the source of text. Furthermore, the coverage of drone images is more limited compared to the OSM and satellite images, making it difficult to achieve large-scale geo-localization.

Table 1. Comparison of the proposed CVG-Text with existing cross-view datasets. # Ref. and # G-Query represent the number of reference images and ground query images, respectively.

Dataset	Text	Pano ¹	Single ²	OSM	# G-Query	# Ref.
CVUSA [38]		✓			44k	44k
CVACT [25]		✓			128k	128k
VIGOR [56]		✓			90k	105k
CVGlobal [45]		✓		✓	130K	130K
Uni.-1652 [53]			✓		14K	90k
Geotext-1652 [9]	✓		✓		14K	90k
Ours	✓	✓	✓	✓	30k	60k

¹ Pano refers to the Panoramic street-view images.

² Single refers to the Single-view street-view images.

3.2. Data Collection

We collected panoramic and single-perspective street-view images from different city areas using the Google Street View ² and Google Places API ³. The resolution of the panoramic street-view images is 2048×1024 , with the north direction aligned to the middle column. The resolution of the single-perspective images is not fixed, but all are high-definition images. Based on the latitude and longitude coordinates of the street-view images, we collected corresponding satellite image tiles using Google Maps API⁴. The size of the satellite images is 512×512 , with a zoom level of 20 and a ground resolution of around 0.12m, aligned with the center of the street-view images.

Additionally, we collected corresponding OSM data for each region in vector format, encompassing global geographic and map information. OSM data contains numerous points of interest (POIs) with rich label information, such as restaurant names and bus stops, which are highly useful for geo-localization tasks. We utilize raster tiles provided by the OSM official website as retrieval targets, retaining various POI identifiers. The size of the OSM raster data is 512×512 , which is close to the image size and corresponding geographical extent of the satellite images. Since OSM data is dynamically updated, we collected street-view data from recent years to minimize discrepancies between the OSM and street-view data. Existing cross-view datasets typically include street-view data from before 2021; thus, rather than adding text annotations to existing datasets, we opted to collect new data.

3.3. Text Data Synthesis

Figure 3 illustrates the main process of text synthesis based on GPT-4o. We first utilize Paddle OCR⁵ to capture text within street-view images, enabling GPT to focus on key semantic text information present in the images while re-

²<https://www.google.com/streetview/>

³<https://developers.google.com/maps/documentation/places/web-service>

⁴<https://www.google.com/maps>

⁵<https://github.com/PaddlePaddle/PaddleOCR>

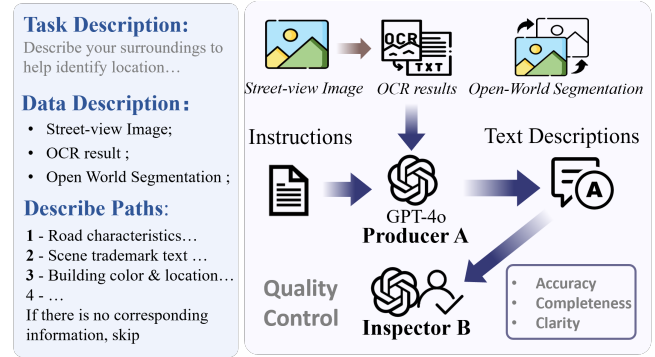


Figure 3. Overall Process for Street-View Text Description Generation using GPT-4o.

ducing hallucination phenomena. Additionally, we perform open-world segmentation on the street-view images to provide more semantic details and location information for the subsequent text synthesis process. With the help of the semantic segmentation results, OCR outputs related to moving objects, such as vehicles, can also be filtered out.

Next, we designed a systematic prompting scheme to generate scene text descriptions using GPT-4o. The system prompts are divided into four parts: task requirements, input data introduction, description paths, and response demands. Specifically, we employ a progressive scene analysis strategy to guide GPT in sequentially describing road features, building signage, and the overall environment, including aspects such as color, material, and distribution. In real-world applications, users may not know the exact geographic orientation. Therefore, text descriptions often use simple directional cues like front, back, left, and right to describe the scene. For more implementation details, refer to the supplementary materials.

In terms of quality control, we first filter out incorrectly formatted responses and conduct GPT-based review of both images and textual descriptions to ensure accuracy, consistency, content completeness, relevance, clarity, and comprehensiveness. Samples that do not meet the standards are re-synthesized. Additionally, we extracted 20% of the samples (approximately 6,000) for manual expert review, involving 10 human evaluators and requiring around 100 hours of work. The pass rate for the manual checks was 77.6%. Furthermore, street-view retrieval experiments using a retrained CLIP model showed that the Top-1 recall rate for text-to-street view was over 85.5%, validating the high quality of the dataset’s text descriptions.

4. Method

4.1. Overview

Problem Formulation. In this work, we introduce a novel task for cross-view geo-localization with natural language. The objective of this task is to leverage natural language

descriptions to retrieve its corresponding OSM or satellite images, of which the location information is usually available. The input for this task is a text description $Q\text{-Text}$ describing a street scene. In practical applications, users often have location prior, such as knowing they are in a particular district of New York City, but not the precise location. Therefore, the retrieval model utilizes this location prior to narrow the search scope M , querying a smaller subset of satellite images, $R\text{-sat}$, and OSM image, $R\text{-OSM}$.

Model Architecture. To address the challenge of the task, we propose the CrossText2Loc architecture (Figure 4). It aligns the image and text domains through an enhanced text embedding module and a contrastive learning loss L_{itc} enabling efficient text retrieval tasks. Furthermore, during inference, we introduce an optional Explainable Retrieval Module (ERM) to provide natural language analysis, enhancing the interpretability and confidence of retrieval decisions.

4.2. Image-text Contrastive Learning

CrossText2Loc adopts a dual-stream architecture, consisting of a text encoder and a visual encoder that extract features from text queries and reference satellite or OSM images, respectively. Since the text describes scenes corresponding to street view images, there is a significant domain gap between these and the reference satellite and OSM data. As a result, the features obtained by the pre-trained encoder are spatially distant. Therefore, we align the image and text embedding spaces through contrastive learning, while jointly training both the text and image encoders. We set the batch size to n and align the representations of images and text by minimizing the contrastive learning loss function L_{itc} . To achieve this, we uniformly encode *panoramic* and *single-view* texts as t , and *satellite* and *OSM* images as v . The loss function is expressed as:

$$L_{itc} = \sum_{i=1}^n \sum_{j=1}^n -\log \frac{\exp(\text{sim}(v_i, t_j)/\tau)}{\sum_{k=1}^n \exp(\text{sim}(v_i, t_k)/\tau)} \quad (1)$$

Where v_i and t_j represent the i -th image and the j -th text embedding vector respectively. τ is a learnable temperature used to control the sharpness of softmax distribution.

Scene description texts are generally complex and lengthy, and their matching with OSM or satellite images is not based on direct surface-level word matching, but requires deeper semantic understanding. However, previous text encoders, such as CLIP[30], have limited positional embeddings, often resulting in truncation and insufficient global text representation. To address this issue, we adopt the **Expanded Positional Embedding (EPE)** method to extend the positional embeddings of the text encoder, which are then fed into the Transformer block. During subsequent embedding in the Transformer architecture, attention mechanisms are used to capture critical localization cues. Specially, we utilize linear interpolation to expand the positional embeddings to accommodate a sequence length of N (300) tokens. The expanded

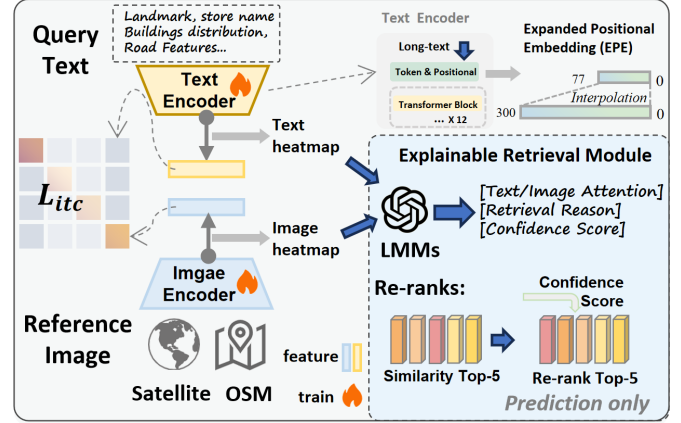


Figure 4. **The proposed CrossText2Loc method.** Street-view texts serve as query inputs, with satellite and OSM images as references.

positional embedding P^* can be computed from the original positional embedding P as follows:

$$P^*(x) = (1 - (x - \lfloor x \rfloor)) \cdot P(\lfloor x \rfloor) + (x - \lfloor x \rfloor) \cdot P(\lceil x \rceil) \quad (2)$$

Where $P^*(x)$ represents the expanded positional embedding at position x , $P(\lfloor x \rfloor)$ and $P(\lceil x \rceil)$ are the values from the original positional embedding at the indices $\lfloor x \rfloor$ and $\lceil x \rceil$, with $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denoting the floor and ceiling functions respectively. Due to the scene texts in our case are generated by GPT for simulating user data, there are no prominent short titles at the beginning which carry significant importance. Therefore, unlike LongCLIP [50] uses Knowledge Preserving Stretching, we use a full interpolation method.

4.3. Explainable Retrieval Module (ERM)

We also propose an optional Explainable Retrieval Module, designed to enhance the interpretability and trustworthiness of geolocation retrieval predictions through natural language. **Attention Heatmap Generation.** Inspired by the method described in [5], we generate attention heatmaps to reveal the model’s focus during retrieval. We initialize the correlation heatmap before the starting layer s with an identity matrix I . Then, we iteratively process each attention layer from the starting layer s to the output layer L . In each attention mechanism, we accumulate non-negative gradient contributions to generate attention heatmaps that highlight the regions the model focuses on. The image or text correlation heatmap R can be calculated as:

$$R^{(l)} = R^{(l-1)} + \frac{1}{H} \sum_{h=1}^H \max(0, \nabla A_h^{(l)} \odot A_h^{(l)}) R^{(l-1)} \quad (3)$$

Where $A_h^{(l)}$ represents the attention weight of the h -th attention head at layer l , $\nabla A_h^{(l)}$ is the gradient of the attention weight, \odot denotes the Hadamard product, H is the total number of attention heads.

LMM Interpretation. After generating the image and text heatmaps, we feed them into the LMM (GPT-4o) with a

Method	Satellite image						OSM data					
	NewYork		Brisbane		Tokyo		NewYork		Brisbane		Tokyo	
	R@1	L@50	R@1	L@50	R@1	L@50	R@1	L@50	R@1	L@50	R@1	L@50
ViLT[19]	11.58	15.58	11.00	14.50	10.83	15.50	5.83	9.92	8.67	11.75	4.67	9.17
X-VLM[48]	15.74	16.86	15.67	17.60	12.46	14.34	16.14	17.26	20.46	21.94	9.53	10.94
SigLIP-B/16[49]	19.67	21.08	19.58	22.00	15.58	17.92	20.17	21.58	25.58	27.42	11.92	13.67
SigLIP-SO400M[49]	33.50	34.83	34.25	36.83	28.42	31.50	27.75	29.58	29.75	31.58	17.50	19.50
EVA2-CLIP-B/16[12]	25.17	26.58	28.42	31.75	22.50	25.25	18.58	20.83	27.33	28.83	13.92	15.67
EVA2-CLIP-L/14[12]	34.08	35.67	35.67	38.00	31.00	34.08	26.33	28.67	30.92	32.67	19.83	22.50
CLIP-B/16[30]	26.67	28.17	29.92	32.58	24.00	27.25	27.42	29.42	30.83	32.67	17.75	19.92
CLIP-L/14[30]	35.08	37.08	34.08	37.25	28.08	30.50	31.50	33.58	32.50	34.67	21.00	23.17
BLIP[22]	34.58	37.25	34.50	38.17	29.75	33.67	52.92	55.92	43.00	46.33	30.67	34.50
Ours (w/o ERM)	<u>46.25</u>	<u>48.75</u>	<u>43.58</u>	<u>47.42</u>	<u>36.83</u>	<u>39.58</u>	<u>59.08</u>	<u>62.00</u>	<u>46.08</u>	<u>48.67</u>	<u>34.33</u>	<u>38.33</u>
Ours	50.33	53.07	47.58	51.80	41.75	43.86	62.33	65.39	48.75	51.50	36.92	41.22

Table 2. **Quantitative comparison of different methods on CVG-Text.** R@1 represents the Top-1 image recall rate; L@50 represents the recall rate where localization error is less than 50 meters. [Key: **Best**, Second Best]

carefully designed prompt. First, the LMM observes the heatmaps to identify key clues in the image and text that the model focuses on during retrieval, such as specific landmarks or geographical features. Then, the LMM compares and reasons over the highlighted regions in the text and image, providing an explanation for why the model retrieved the query in a particular way, i.e., the rationale behind the matching of the query and reference data. Finally, the LMM outputs the confidence level of the retrieval rationale, simulating user decision-making. LMMs are capable of providing retrieval explanations in natural language, beyond just similarity measures, which is highly beneficial for the interpretability and trustworthiness of geo-localization retrieval.

Confidence re-ranking. The confidence scores obtained through explainable retrieval can be used to re-rank the original top-5 similarity scores from the retrieval match. We set the results with top-1 confidence scores lower than 0.5 and normalize both the similarity and confidence scores to the [0, 1] range, summing them to obtain the re-ranked Top-5 results. In the subsequent experimental section, we demonstrate that the effective re-ranking strategy helps simulate user retrieval decisions and improves recall in tasks.

5. Experiment

5.1. Experimental Setup

Implementation Details. We used the CLIP-L/14@336px model [30] pre-trained by OpenAI as the backbone, with the Adam optimizer [20], a learning rate of $1e-5$, and cosine learning rate decay. The batch size was set to 128, and training was conducted over 40 epochs on four NVIDIA A100 GPUs. The image resolution was set to default 336×336 , and the text context length N was 300 tokens. We initialized the temperature coefficient τ from the checkpoint.

Moreover, as mentioned in Section 4.1, based on practical application requirements and utilizing user location priors, we set the retrieval range M to 100. During testing, the reference database consists of 100 samples from the nearby area, covering an area of approximately 10 km². M is a configurable parameter, and we present additional results with different M settings in the supplementary material. The image resolution for all comparison methods follows their respective default best settings, and all methods are trained on the same proposed dataset.

Evaluation Metrics. Following previous cross-view geo-localization works [10, 33], we use the image recall accuracy of the top K images as an evaluation metric to assess text retrieval localization performance. Specifically, given a query text for a certain location, if its ground-truth OSM and satellite image is within the top K retrieval results, the location is considered “successfully localized.” Additionally, we also provide the localization recall rate metric [21, 40]. Similarly, the localization recall rate refers to the proportion of retrieved results where the distance to the actual location is below a specified threshold.

5.2. Geo-localization Performance

We evaluated the performance of various text-based retrieval methods under different settings of satellite images and OSM data, with the results shown in Table 2. Among the existing approaches, BILP achieves optimal performance, as it is not constrained by limitations on text embedding length. Our method, even without the ERM, achieved the best results. Compared to the baseline CLIP method, the proposed approach improved Recall@1 by 14.1% and Recall@10 by 14.8%, demonstrating its advantages in this task.

Next, we evaluated the performance of methods in different cities and across different data sources (Satellite/OSM).

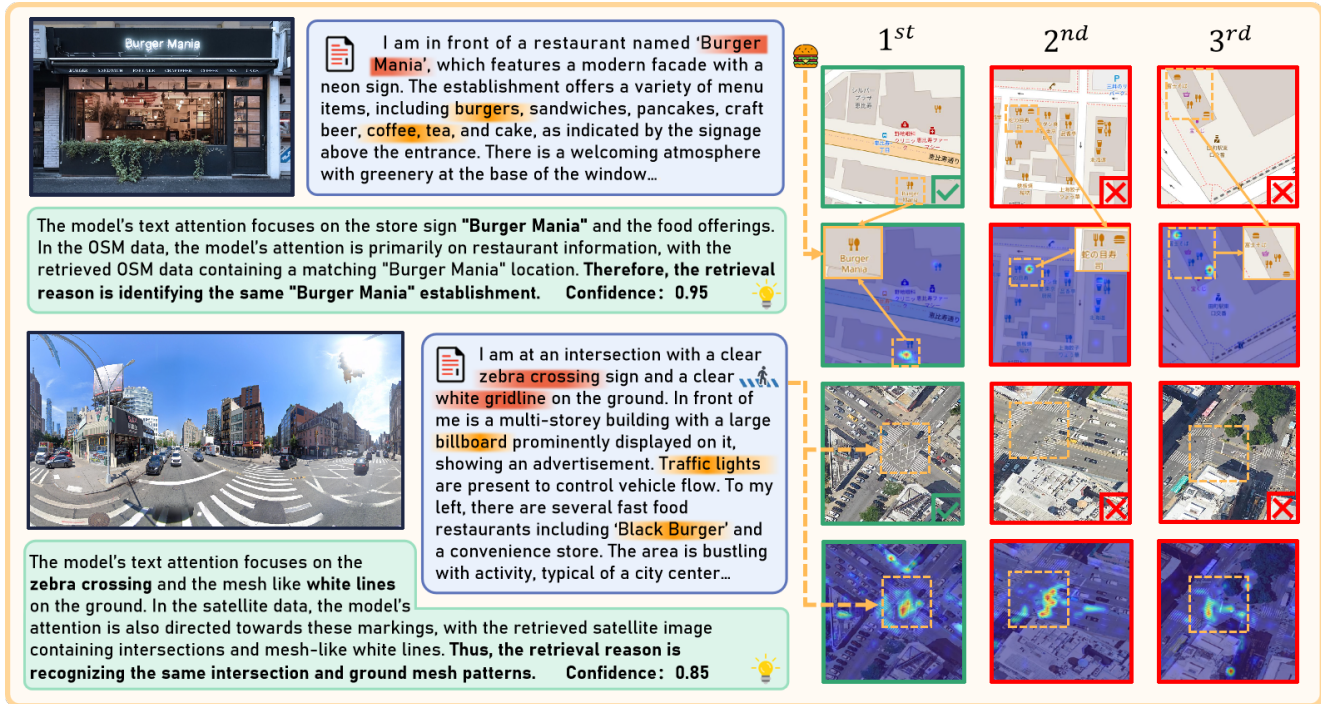


Figure 5. **Qualitative retrieval results on CVG-Text Dataset.** The left side of the figure displays the original street-view data, synthetic text data with corresponding response heatmaps, and retrieval reason provided by our ERM module. The right side shows the top three retrieval results with corresponding response heatmaps; green indicates correct matches and red denotes incorrect results.

In New York, OSM outperformed satellite imagery due to richer POI data, such as bus stops and store names, which are hard to identify in satellite images. In contrast, the level of detail in OSM data for Brisbane is limited, and in this case, the localization performance of satellite images is comparable to that of OSM. In Tokyo, due to the poor pre-training of CLIP on Japanese, the model’s response to certain Japanese words in the street view description text and Japanese POIs in OSM data is weak, leading to the least favorable performance in Tokyo. The supplementary materials also include cross-city evaluations and a collaborative retrieval method for OSM and satellite images.

5.3. Explainable Retrieval

We present the results of the Explainable Retrieval Module (ERM) in Figure 5. This method uses text and the heatmap responses of the retrieved images to highlight the key features the localization model focuses on. In the first example, the model focuses on “Burger Mania”. Interestingly, even in non-top-1 results, it still emphasizes the burger icon. In the second example, the model focuses on the “zebra crossing” and “white gridline”, with the first three retrieval results all highlighting the zebra crossing, and the best retrieval result matching both features. We provide a quantitative evaluation of the ERM module’s retrieval rationale in the supplementary materials, using similarity matching with human-written

explanations, CIDEr[36], ROUGE-L[24], and multidimensional human scoring.

Moreover, the confidence scores obtained from LMMs essentially simulate the user’s confidence decision-making process. When the confidence score is low, it indicates that, even though the similarity score of the result may be high, it is still not convincing. In this case, by applying the re-ranking strategy, we can obtain better retrieval results, as shown in Table 2. The practical value of the Explainable Retrieval Module (ERM) lies in assisting users to assess the rationale behind the localization decision. They can select the most reasonable result from similar candidates, whereas previous methods only relied on the highest similarity match. If the rationale provided by ERM is not convincing enough, users can choose to re-rank the remaining search results or add additional descriptions to provide more visual clues.

Additionally, we generate feature heatmap responses based on text and reference images. By analyzing the top-ranked words in the model’s attention, we identified eight focal subcategories, such as “LandMarker” for landmarks and “SignName” for store names or bus stop names. A list of these words and categories is in the supplementary materials. Attention scores for each category in the scene are shown in Figure 6. Results reveal that the cross-view retrieval model mainly focuses on landmarks and road information, with less attention to vehicles, sky, and weather, as

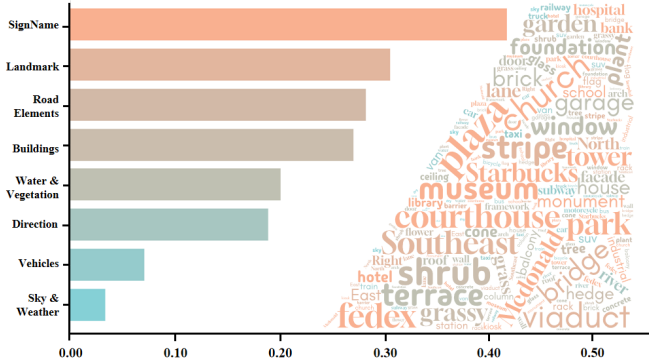


Figure 6. The average attention score for each object category (left) and the word cloud of selected attention words (right).

Methods	Text quality			Retrieval result	
	Len	TTR	Simi.	R@1-OSM	R@1-Sat
Baseline	108	0.74	0.22	25.17	38.00
Ours	126	0.76	0.17	59.08	46.25

Table 3. Comparison between Baseline (directly using GPT-generated Text) and our CVG-Text dataset in New York.

Method	Satellite		OSM	
	R@1	R@5	R@1	R@5
SigLIP	19.67	53.33	20.17	41.83
SigLIP + EPE	29.50	64.00	45.25	70.25
CLIP	35.08	71.42	31.50	55.42
CLIP + EPE	46.25	81.58	59.08	82.75

Table 4. Ablation Study of Expanded Positional Embedding (EPE) Module in New York.

they are less relevant for localization. This highlights key clues for understanding cross-view retrieval models.

5.4. Ablation Studies

We evaluated the text synthetic effects of those directly generated by GPT versus those generated through the integrated process of CVG-Text, as shown in Table 3. CVG-Text exhibits higher text length and vocabulary complexity (TTR), with lower text similarity, indicating better text quality. The inclusion of OCR assistance helps precisely capture textual details in street view images, reducing GPT’s tendency for vague descriptions and hallucinations, and effectively generates text with key localization cues, significantly improving OSM retrieval performance. Open-World Segmentation further enhances GPT’s semantic and spatial understanding, improving satellite image retrieval performance.

We employed two different learning architectures, CLIP [30] and SigLIP [49], to validate the contribution of our proposed Expand Positional Embedding (EPE) module. As

Method	OSM		Satellite		R@1
	R@1	R@10	R@1	R@10	average
SAFA [33]	19.25	44.88	77.40	95.30	48.32
Geo-Dtr [51]	24.10	53.30	86.45	98.80	55.28
Sample4G [10]	27.10	62.90	<u>91.70</u>	<u>99.20</u>	59.40
Text-only	<u>59.08</u>	<u>90.00</u>	46.25	91.00	52.67
Sample4G [10] +Text	67.30	97.20	98.40	99.80	82.80

Table 5. Impact of Adding Text Branch for Cross-View Retrieval.

shown in Table 4, the use of the EPE module significantly improves the performance of both text retrieval methods on satellite images and OSM, with R@1 recall rates increasing by 10.5% and 26.3%, respectively. This demonstrates the effectiveness of extending text encoding length for handling longer text content in this task.

5.5. Further Discussion on Cross-view retrieval task

Previous cross-view retrieval tasks have primarily focused on image-to-image queries. Leveraging the CVG-Text dataset, we expanded this task to multimodal image-text joint queries. Building on the state-of-the-art cross-view street-view retrieval method, Sample4G [10], we added an additional text query branch via fusing the similarity scores. We conducted satellite and OSM retrieval experiments in the New York area of the CVG-Text dataset. As shown in Table 5, compared to using only street-view images as queries, the addition of a text branch improved Top-1 recall by 6.7% for the satellite retrieval task and by 40.2% for the OSM retrieval task, effectively enhancing retrieval accuracy. Since the detailed scene descriptions in CVG-Text align more closely with POI data in OSM, the improvement for OSM retrieval is more pronounced than for satellite image data.

6. Conclusion

In this work, we explore the task of cross-view geolocalization using natural language descriptions and introduce the CVG-Text dataset, which includes well-aligned street-views, satellite images, OSM images, and text descriptions. We also propose the CrossText2Loc text retrieval localization method, which excels in handling long-text retrieval and interpretability for this task. This work represents another advancement in the field of natural language-based localization. It also introduces new application scenarios for cross-view localization, encouraging subsequent researchers to explore and innovate further.

Acknowledgements

This work was supported in part by the Natural Science Foundation of Guangdong Province, China (Grant No. 2025A1515010400), the National Natural Science Foundation of China (Grant No. 42201358) and Shanghai Artificial Intelligence Laboratory.

References

- [1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10): 105–112, 2011. 1
- [2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 2
- [3] Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. Accessed: 2024-09-23. 3
- [4] Anh-Quan Cao, Maximilian Jaritz, Matthieu Guillaumin, Raoul de Charette, and Loris Bazzani. Latteclip: Unsupervised clip fine-tuning via Imm-synthetic texts. *arXiv preprint arXiv:2410.08211*, 2024. 3
- [5] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021. 5
- [6] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019. 2
- [7] Jiaqi Chen, Daniel Barath, Iro Armeni, Marc Pollefeys, and Hermann Blum. ” where am i?” scene retrieval with language. *arXiv preprint arXiv:2404.14565*, 2024. 1
- [8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 2
- [9] Meng Chu, Zhedong Zheng, Wei Ji, and Tat-Seng Chua. Towards natural language-guided drones: Geotext-1652 benchmark with spatially relation matching. *arXiv preprint arXiv:2311.12751*, 2023. 3, 4
- [10] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16847–16856, 2023. 1, 2, 6, 8
- [11] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 3
- [12] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. 6
- [13] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building rome on a cloudless day. In *Computer Vision—ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part IV 11*, pages 368–381. Springer, 2010. 1
- [14] Owen He, Ansh Jain, Axel Adonai Rodriguez-Leon, Arnav Taduvayi, and Matthew Louis Mauriello. From crowdsourcing to large multimodal models: Toward enhancing image data annotation with gpt-4v. 2023. 2, 3
- [15] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023. 2
- [16] Zilong Huang, Jun He, Junyan Ye, Lihan Jiang, Weijia Li, Yiping Chen, and Ting Han. Scene4u: Hierarchical layered 3d scene reconstruction from single panoramic image for your immerse exploration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26723–26733, 2025. 1
- [17] Arnold Irschara, Christopher Zach, Jan-Michael Frahm, and Horst Bischof. From structure-from-motion point clouds to fast location recognition. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2599–2606. IEEE, 2009. 1
- [18] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 9(2):1286–1293, 2023. 2
- [19] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 6
- [20] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [21] Manuel Kolmet, Qunjie Zhou, Aljoša Ošep, and Laura Leal-Taixé. Text2pos: Text-to-point-cloud cross-modal localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6687–6696, 2022. 1, 2, 6
- [22] 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. 6
- [23] Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Song Xia, and Dahua Lin. Omniscity: Omnipotent city understanding with multi-level and multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17397–17407, 2023. 2
- [24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 7
- [25] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of*

- the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019. 4
- [26] Simon Lynen, Bernhard Zeisl, Dror Aiger, Michael Bosse, Joel Hesch, Marc Pollefeys, Roland Siegwart, and Torsten Sattler. Large-scale, real-time visual–inertial localization revisited. *The International Journal of Robotics Research*, 39(9):1061–1084, 2020. 1
- [27] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. Openmvg: Open multiple view geometry. In *Reproducible Research in Pattern Recognition: First International Workshop, RRRP 2016, Cancún, Mexico, December 4, 2016, Revised Selected Papers 1*, pages 60–74. Springer, 2017. 1
- [28] OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. 2, 3
- [29] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Philip Torr, Zhe Lin, and Jiaya Jia. Open world entity segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8743–8756, 2022. 2
- [30] 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. 3, 5, 6, 8
- [31] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023. 2, 3
- [32] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12716–12725, 2019. 1
- [33] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019. 6, 8
- [34] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020. 2
- [35] Huilin Tian, Jingke Meng, Wei-Shi Zheng, Yuan-Ming Li, Junkai Yan, and Yunong Zhang. Loc4plan: Locating before planning for outdoor vision and language navigation. *arXiv preprint arXiv:2408.05090*, 2024. 1, 2
- [36] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 7
- [37] Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *Advances in Neural Information Processing Systems*, 36, 2024. 1
- [38] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocation with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015. 3, 4
- [39] Yan Xia, Yusheng Xu, Cheng Wang, and Uwe Stilla. Vpc-net: Completion of 3d vehicles from mls point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174:166–181, 2021. 1
- [40] Yan Xia, Letian Shi, Zifeng Ding, Joao F Henriques, and Daniel Cremers. Text2loc: 3d point cloud localization from natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14958–14967, 2024. 1, 2, 6
- [41] Zimin Xia, Yujiao Shi, Hongdong Li, and Julian FP Kooij. Adapting fine-grained cross-view localization to areas without fine ground truth. In *European Conference on Computer Vision*, pages 397–415. Springer, 2025. 2
- [42] Zhiyuan Yan, Junyan Ye, Weijia Li, Zilong Huang, Shenghai Yuan, Xiangyang He, Kaiqing Lin, Jun He, Conghui He, and Li Yuan. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*, 2025. 3
- [43] 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. 3
- [44] Junyan Ye, Qiyan Luo, Jinhua Yu, Huaping Zhong, Zhimeng Zheng, Conghui He, and Weijia Li. Sg-bev: Satellite-guided bev fusion for cross-view semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27748–27757, 2024. 1
- [45] Junyan Ye, Zhutao Lv, Weijia Li, Jinhua Yu, Haote Yang, Huaping Zhong, and Conghui He. Cross-view image geo-localization with panorama-bev co-retrieval network. *arXiv preprint arXiv:2408.05475*, 2024. 1, 2, 4
- [46] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *arXiv preprint arXiv:2410.09732*, 2024. 2
- [47] Junyan Ye, Jun He, Xiang Zhang, Yi Lin, Honglin Lin, Conghui He, and Weijia Li. Satellite image synthesis from street view with fine-grained spatial textual guidance: A novel framework. *IEEE Geoscience and Remote Sensing Magazine*, 2025. 3
- [48] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, pages 25994–26009. PMLR, 2022. 3, 6
- [49] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 6, 8
- [50] 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, 2025. 3, 5
- [51] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceed-*

- ings of the AAAI Conference on Artificial Intelligence*, pages 3480–3488, 2023. [8](#)
- [52] Junwei Zheng, Ruiping Liu, Yufan Chen, Kunyu Peng, Chengzhi Wu, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. Open panoramic segmentation. In *European Conference on Computer Vision*, pages 164–182. Springer, 2025. [2](#)
- [53] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1395–1403, 2020. [4](#)
- [54] 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. [3](#)
- [55] Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10707–10715, 2025. [3](#)
- [56] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. [2](#), [3](#), [4](#)