Trajectory-aware Cross-view Geo-localization with Sequential Observations

Tianyi Gao, Jiayu Lin, Danielle Beaulieu, and Nathan Jacobs

Washington University in St. Louis, St. Louis, MO 63130, USA
{t.gao, jiayu.lin, b.danni, jacobsn}@wustl.edu

Abstract. Cross-view geo-localization matches ground-level observations against geo-tagged satellite imagery. Recent methods show that sequential queries such as video clips yield richer spatiotemporal cues than single images, yet they overlook a complementary sequential modality: route descriptions—which capture the same trajectory at a higher level of abstraction and are often the only input available (e.g., a user directing an autonomous vehicle to a pickup point). To bridge this gap, we introduce SeqGeo-VL, a dataset of ∼39K video–text–satellite triplets, and TrajLoc, a unified framework capable of processing both video clips and route descriptions. By leveraging both dense visual and abstract linguistic semantics, TrajLoc enables these modalities to mutually reinforce cross-view matching. We further propose TrajMod, a lightweight module that conditions query embeddings on trajectory geometry, yielding spatially-aware representations. Experiments show that TrajLoc achieves substantial gains over state-of-the-art methods on both video and text geo-localization. Code, model weights, and the dataset are released at https://humblegamer.github.io/trajloc/.

1 Introduction

Urban embodied AI agents, such as autonomous vehicles and quadrupeds, are increasingly deployed for city-wide services [12]. Yet, in crowded urban environments, GPS signals are frequently noisy or completely degraded by tall buildings and vegetation, rendering global localization unreliable [27,32]. As a result, precise spatial localization based on egocentric visual observations or natural-language descriptions has emerged as a fundamental necessity for agents to navigate safely and cooperate with humans [1,20], unlocking applications like pedestrian assistance, goods delivery [27], and vehicle pickup [9].

Cross-view geo-localization addresses this challenge by matching ground images or textual descriptions against a GPS-tagged satellite image database, casting localization as a retrieval problem [10, 26, 40]. Although its granularity is bounded by satellite-patch resolution, it meets the needs of most tasks [32] while offering clear advantages in coverage and storage cost, making it a promising avenue for scalable localization [32].

To overcome perceptual aliasing—where visually identical structures like uniform facades or repeating city blocks cause catastrophic matching ambigui-

ties—recent works [16, 19, 21, 25] use sequential observations as queries (as illustrated in Fig. 1). Sequences provide richer cues (motion continuity, changing viewpoints, route-level semantics) that significantly improve retrieval robustness over isolated frames. While this video-to-image retrieval formulation provides dense visual evidence to disambiguate places, it overlooks the textual narrative modality. This limits applications when visual inputs are unavailable [29] or during human-robot collaboration, where users naturally rely on abstract language rich in spatiotemporal cues.

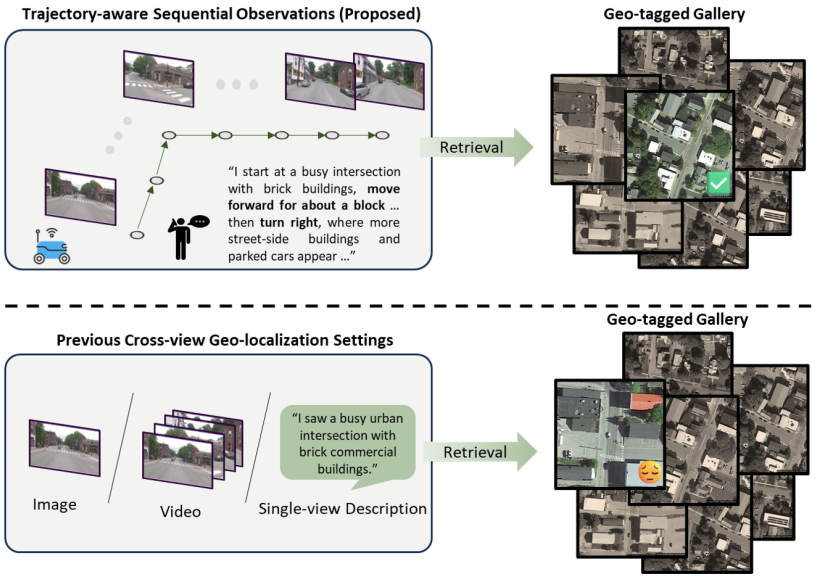

Fig. 1: Motivation for our approach, which extends existing cross-view geo-localization methods to sequential route descriptions and explicitly incorporates trajectory information into video/text queries, producing spatially-grounded representations.

However, existing cross-view geo-localization datasets lack route descriptions, let alone video–text–satellite triplets. This restricts flexibility and leaves the potential synergy between route descriptions and videos of a trajectory unexplored.

To address this gap, we develop a scalable annotation pipeline—powered by open-source VLMs and LLMs—that simulates how human users describe routes while moving. Using a progressive strategy, this pipeline combines trajectory-shape priors, frame-level VLM labeling, LLM summarization for concise narratives, and sample-based human verification. We apply this pipeline to the widely used SeqGeo dataset to create **SeqGeo-VL**, a novel multimodal benchmark comprising 38,863 triplets of trajectory videos, route descriptions, and corresponding geo-referenced satellite images.

Building on SeqGeo-VL, we propose **TrajLoc**, a unified approach that supports cross-view geo-localization using either video queries or route descriptions. We employ a shared satellite image encoder alongside two separate encoders to process video and route descriptions, respectively. Since directly training modality alignments of varying difficulties can hinder the optimization of the shared satellite image encoder, we adopt a two-stage curriculum learning recipe to maximize the benefits of co-training.

Furthermore, we observe that general CLIP and MLLM embeddings often fail to capture spatial layout, a capability that is crucial for geo-localization since semantically similar objects in different spatial arrangements correspond to distinct locations. We advocate addressing this by better utilizing trajectory information. We leverage the fact that modern mobile devices are inherently equipped with IMUs and compasses capable of capturing trajectory geometry alongside video. While often stripped from standard web videos, this information is readily available in active navigation scenarios yet frequently overlooked in cross-view geo-localization methods. To this end, we introduce **Trajectory-conditioned Modulation (TrajMod)**, a lightweight plug-and-play module that conditions on trajectory geometry to modulate query embeddings, yielding spatially grounded representations and improving cross-view matching robustness. Our main contributions are:

- **Task and Dataset:** We present a progressive VLM-LLM annotation pipeline to ensure high-quality descriptions and introduce **SeqGeo-VL**, a multi-modal dataset featuring $38,863$ aligned video–text–satellite triplets for evaluating both video- and text-based geo-localization.
- **Unified Approach:** We propose **TrajLoc**, a strong baseline that handles both video and route description queries via a two-stage curriculum learning strategy, unlocking cross-modal co-training gains.
- **Spatially Grounded Representation:** We introduce **TrajMod** to inject trajectory geometry into vision-language embeddings. This produces spatially grounded representations that achieve state-of-the-art performance on both cross-view video and text geo-localization tasks.

2 Related Work

2.1 Cross-view Geo-localization with Video

Recent progress in cross-view geo-localization has extended ground-to-aerial image matching to sequential settings with video queries. Many datasets formulate cross-view video geo-localization as a retrieval task, including GaMa [21] (based on BDD100K [34]), SeqGeo [38], and CVLNet [19] (based on KITTI [7]). While these datasets facilitate learning from sequential observations, they primarily support video-based queries and do not provide natural language descriptions of trajectories, limiting their applicability to scenarios where only route descriptions are available. To bridge this gap, we build upon SeqGeo [38] and introduce

a progressive annotation pipeline to enrich it with trajectory-level textual descriptions, forming a benchmark for cross-view geo-localization with multimodal sequential observations.

Building on these datasets, recent methods focus on mining video context to improve robustness under large viewpoint and appearance changes [16,21,25,38]. Focusing on sophisticated aggregation mechanisms for sequential observations, GaMa [21] utilizes a 3D-CNN, SeqGeo [38] employs a Temporal Feature Aggregation Module for mixing, GARet [16] introduces a transformer-based adapter for fusion, and FlexGeo [25] uses inter-frame similarity as guidance. Existing approaches rarely exploit explicit trajectory geometry, like per-frame headings or origin–destination layout, to inform representation learning. Furthermore, existing methods typically assume sequential observations restricted to video format; however, leveraging long-form route descriptions—a scenario of immense practical value for human-robot interaction—remains underexplored.

2.2 Cross-view Geo-localization with Natural Language Description

There is abundant work on retrieval tasks between image and text modalities [18, 23, 28], especially using contrastive learning [17]. However, limited work exists in the domain of cross-view geo-localization via natural language. The goal of this domain is to retrieve the corresponding satellite image from a visually grounded text description. Identifying an image from text that is an exact match to the intended geolocation is an especially challenging task; text often lacks the fine detail required to distinguish one location from another visually similar location.

The primary prior work pursuing this task, CrossText2Loc [32], utilizes additional contrastive learning on top of a CLIP backbone to align the embedding spaces of text descriptions and satellite images. They also propose an Expanded Positional Embedding, which accommodates more details in the descriptions. As a pioneering contribution, they constructed the CVG-Text dataset and established an effective VLM-based annotation pipeline. However, since their method focuses on describing single isolated street-view images, it does not yet account for the multiple or sequential observations inherent in long-form trajectory descriptions. Motivated by this, we design a progressive annotation pipeline powered by the synergy of VLM perception and LLM reasoning, applied to video sequences with trajectory metadata.

2.3 Spatially Grounded Vision-Language Embedding

Recent research [35] suggests that CLIP behaves as "bag-of-words" models, lacking a fundamental comprehension of spatial relationships. Although CLIP exhibits strong zero-shot capabilities across diverse tasks, it consistently struggles with understanding the spatial arrangement of objects within images [22]. Wang et al. [22] demonstrate that CLIP often fails to distinguish correct spatial descriptions from incorrect ones, sometimes even assigning higher matching scores to captions with erroneous spatial relationships. Even Multi-modal Large Language Models (MLLMs) with sophisticated decoders struggle to handle relative

directions or spatial layouts. The VSI-Bench [31] highlights that even strong MLLMs with advanced video understanding capabilities cannot accurately map egocentric video frames to allocentric object positions and camera trajectories, an area that remains under active development.

To develop spatially grounded representations, recent studies like SpatialRGPT [4] and SpatialCLIP [22] incorporate depth information with RGB imagery. This approach is designed to bolster the model's grasp of basic spatial properties (e.g., size and length) as well as more intricate relationships, including directionality, spatial adjacency, and perspective transformation.

Towards larger-scale outdoor scenarios, some methods integrate structural data; for instance, UGE [37] embeds a spatial topology, which is built upon street-view imagery, road networks, and Point of Interest (POI) data, into the VLM using a graph representation. Although this facilitates high-accuracy geo-localization, our approach operates under weaker assumptions regarding available metadata, requiring neither POI data nor road networks.

3 Problem Formulation and Benchmark

We first formalize cross-view geo-localization from sequential observations and then describe the construction of our multimodal benchmark for this task.

3.1 Task Definition

We study two instances of cross-view geo-localization from sequential observations. A trajectory τ is modeled as a sequence of waypoints with relative coordinates $(\Delta x_i, \Delta y_i)$ and headings h_i. Along τ, sequential observations $O_{1:N}$ can be provided as a video V (detailed) or a route description T (abstract).

$$\tau = \{(\Delta x_i, \Delta y_i, h_i)\}_{i=1}^{N} \quad O_{1:N} \in \{V, T\} \tag{1}$$

Given a geo-tagged satellite gallery $\mathcal{S} = \{S_j\}_{j=1}^{M}$, the goal is to retrieve the matching satellite image:

$$j^* = \arg \max_{j \in \{1,\dots,M\}} \text{sim}(\phi_q(O_{1:N}, \tau),\, \phi_s(S_j))\,, \tag{2}$$

where ϕ_q is a video/text query encoder, ϕ_s is a satellite encoder, and $\text{sim}(\cdot,\cdot)$ denotes similarity measurement. Incorporating τ enables a trajectory-aware cross-view geo-localization formulation, while remaining an optional component contingent on trajectory availability. The agent's location is subsequently derived from the geo-tag associated with the retrieved satellite image.

3.2 Benchmark Dataset

As shown in Table 1, most existing geo-localization datasets rely on single-image queries. While a few explore sequential video queries or single-text queries, none

investigate the sequential route description. To address this gap, we extend the SeqGeo dataset [38] to create SeqGeo-VL, a multimodal geo-localization dataset featuring sequential observations.

To produce high-quality trajectory-level descriptions, we employ a hierarchical annotation pipeline consisting of four stages: (1) **Semantic Extraction**: We utilize Qwen3-VL-8B [2] to generate detailed captions for individual frames, capturing fine-grained visual landmarks. (2) **Motion Integration**: We augment these captions with explicit motion primitives (e.g., *turning left*, *proceeding forward*) derived from trajectory metadata. (3) **Structural Summarization**: Qwen3-30B [30] distills the comprehensive visual semantics and motion cues into a coherent, trajectory-level narrative. This approach ensures that SeqGeo-VL encapsulates both frame-specific details and the global temporal context of the route. (4) **Quality Control**: Three trained annotators scored a geographically uniform ∼1% subset (400 trajectories) on 3-point object fidelity (OF), 3-point spatiotemporal consistency (ST), and route completeness (RC), with inter-annotator agreement measured by Gwet's AC2. A sample passes if RC is rated complete by a majority and $\mathrm{OF} + \mathrm{ST} > 3$, giving a pass rate of 91.8%. Retrieval performance and word statistics provide further indirect measurement.

The dataset comprises 38,863 triplets consisting of videos, route descriptions, and geo-tagged satellite images at zoom level 20. The distribution of extracted keywords is visualized in Fig. 2, and the entire annotation pipeline is summarized in Fig. 3. Following the protocol of the original SeqGeo [38], we split the data into training and testing sets using an 80:20 ratio.

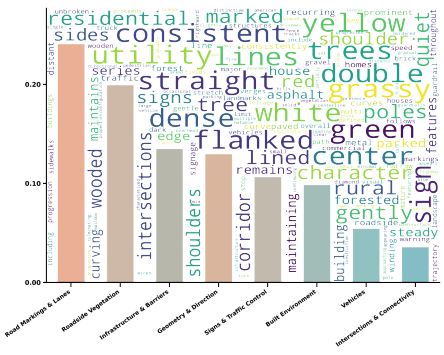

Fig. 2: Word cloud and distribution of the extracted keywords within the SeqGeo-VL dataset.

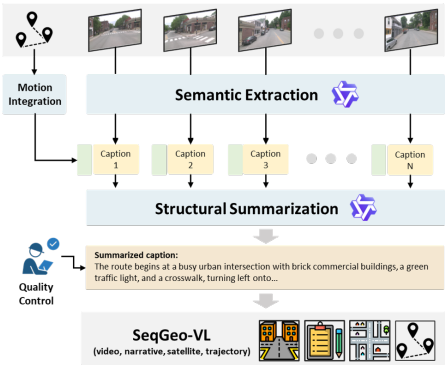

Fig. 3: The proposed hierarchical annotation pipeline for SeqGeo-VL.

Table 1: Comparison of cross-view datasets. **Text** and **Sequential** indicate whether the dataset supports text modalities and sequential observations, respectively. **View Type** describes the field of view of the images. **Qry** and **Ref** denote the total number of query and reference images. **Obs** denotes the average number of observations (i.e., frames) associated with each query instance. † denotes Qry as reported in [25].

Dataset	Text	Sequential	View Type	Qry	Ref	Obs
CVUSA [24]	✗	✗	Panoramic	∼1.5M	∼1.5M	∼1
CVACT [11]	✗	✗	Panoramic	137K	137K	∼1
VIGOR [41]	✗	✗	Panoramic	105K	90K	∼1
CVGlobal [33]	✗	✗	Panoramic	134K	134K	∼1
University-1652 [39]	✗	✗	Limited FOV	5.5K	90K	∼1
GeoText-1652 [5]	✓	✗	Limited FOV	14K	90K	∼1
CVG-Text [32]	✓	✗	Limited FOV	30K	60K	∼1
KITTI-CVL [19]	✗	✓	Limited FOV	28K	41K	∼4
GAMa [21]	✗	✓	Limited FOV	∼15M†	1.9M	–
SeqGeo [38]	✗	✓	Limited FOV	118K	38K	∼7
SetVL480K [25]	✗	✗	Limited FOV	480K	16K	∼4
Ours (SeqGeo-VL)	✓	✓	Limited FOV	118K	38K	∼7

4 Approach

We propose TrajLoc, a unified framework for trajectory-conditioned cross-view geo-localization that supports both video queries and textual route descriptions (Fig. 4). The architecture comprises three CLIP-initialized encoders—for video, text, and satellite imagery—that are trained with a two-stage curriculum. In stage one, we optimize a video–satellite contrastive objective, providing a strong initialization of the satellite encoder for ground-view semantics. In stage two, we freeze the video encoder and align the text encoder to the satellite encoder with a contrastive loss, applying cosine-similarity regularization to prevent the satellite encoder from drifting.

To facilitate allocentric reasoning, we derive a trajectory geometry embedding from waypoint headings and OD bearing. Conditioned on this, TrajMod predicts scale and shift parameters that modulate the video or text query embeddings, yielding spatially-aware representations for cross-view retrieval.

4.1 Encoder Adaptation

We argue that learning text-to-satellite and video-to-satellite retrieval can be mutually beneficial: route descriptions provide abstract, structured semantics, while videos offer dense visual evidence. A unified model also reduces parameters compared to training separate systems. To this end, we instantiate a three-encoder architecture initialized from pretrained CLIP [17].

Video encoder: Given a ground-level video $V = \{v_1, v_2, \ldots, v_N\}$ consisting of N frames sampled along a trajectory, we independently encode each frame using the CLIP ViT-L/14 image backbone to obtain frame-level features

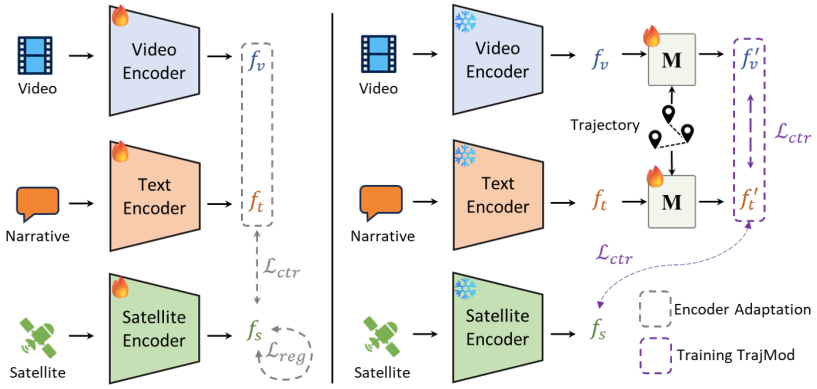

Fig. 4: Overview of the TrajLoc framework. **Left**: Encoder adaptation with a two-stage curriculum. Three CLIP-initialized encoders produce video ($\mathbf{f}_v$), text ($\mathbf{f}_t$), and satellite ($\mathbf{f}_s$) embeddings. Stage 1 trains the video and satellite encoders with a contrastive loss $\mathcal{L}_{\text{ctr}}$; Stage 2 freezes the video encoder and aligns the text encoder, with a regularization loss $\mathcal{L}_{\text{reg}}$ to prevent satellite-encoder drift. **Right**: All three encoders are frozen, and TrajMod (**M**) conditions the video and text embeddings on trajectory geometry to produce spatially grounded representations $\mathbf{f}'_v$ and $\mathbf{f}'_t$ for cross-view retrieval.

$\{\mathbf{f}_{v_1}, \ldots, \mathbf{f}_{v_N}\}$. Following prior video–language models [3,13], we aggregate these into a single video embedding via mean pooling: $\mathbf{f}_v = \frac{1}{N}\sum_{i=1}^{N} \mathbf{f}_{v_i}$. Although there are more sophisticated temporal aggregation strategies (e.g. temporal transformers [38]), we find that mean pooling provides a strong and efficient baseline, keeping the architecture lightweight; improvements can be attributed to the training strategy and trajectory-conditioned modulation rather than encoder complexity.

Text encoder: Route descriptions in SeqGeo-VL are substantially longer than typical CLIP captions (averaging ∼129 tokens), exceeding the default CLIP context window of $L_0 = 77$ tokens. Following CrossText2Loc [32], we initialize the text encoder from OpenAI CLIP and extend its context capacity by linearly interpolating the positional embeddings. This preserves the pretrained positional semantics while smoothly extending the context window, enabling the encoder to process long trajectory-level narratives.

Satellite encoder: We use a CLIP ViT-L/14 image encoder to embed each geo-referenced satellite patch into the shared retrieval space. Because the satellite encoder serves as the common reference for both video and text queries, it plays a central role in the framework. Weights are shared across both training stages and are regularized during the second stage (Sec. 4.3) to prevent catastrophic drift, ensuring that the retrieval space remains coherent for both modalities.

4.2 Trajectory-conditioned Modulation

Rather than reducing cross-view geo-localization to standard appearance-based retrieval, we explicitly incorporate trajectory geometry—a widely available yet historically overlooked modality—to spatially ground our representations.

Trajectory geometry embedding. We use two complementary cues (relative to True North): (i) waypoint heading angle $\mathbf{h} = \{h_1, \ldots, h_T\}$, capturing local turning patterns, and (ii) the OD bearing angle θ_{OD}, providing a stable global orientation. These are lightweight to obtain and do not require metric-accurate geometry. To encode angular variation, we apply Fourier features. For any angle $\alpha \in \mathbf{h} \cup \{\theta_{\text{OD}}\}$,

$$\Phi(\alpha) = \big[\sin(2^0\pi\alpha), \cos(2^0\pi\alpha), \ldots, \sin(2^{K-1}\pi\alpha), \cos(2^{K-1}\pi\alpha)\big], \tag{3}$$

and concatenate the encoded headings and OD bearing along channel dimensions to form a trajectory geometry embedding $\mathbf{z}_{\text{traj}}$.

Spatially-grounded modulation. Inspired by FiLM [15], TrajMod conditions query embeddings on $\mathbf{z}_{\text{traj}}$ via a multi-layer perceptron (MLP) denoted by MLP_m. Given video and text embeddings $\mathbf{f}_v$ and $\mathbf{f}_t$, we learn modality-specific modulation parameters and transform the features where $\odot$ is element-wise multiplication. We use separate MLPs for video and text to account for the modality gap.

$$\boldsymbol{\gamma}_m, \boldsymbol{\beta}_m = \text{MLP}_m(\mathbf{z}_{\text{traj}}), \quad m \in \{v, t\}, \quad \mathbf{f}'_m = \boldsymbol{\gamma}_m \odot \mathbf{f}_m + \boldsymbol{\beta}_m \tag{4}$$

4.3 Training

Aligning route descriptions with satellite images is typically harder than aligning videos with satellite images due to the severe modality gap between abstract text and dense overhead pixels. We therefore train TrajLoc with a curriculum to improve both tasks. In the first stage, we optimize an InfoNCE objective [14] $\mathcal{L}_{\text{ctr}}$ to align video and satellite embeddings, yielding a shared space that retains semantics from both views. In the second stage, we freeze the video encoder and use the satellite encoder as a soft anchor: we align text and satellite embeddings with InfoNCE while regularizing the satellite encoder with a cosine-similarity constraint to mitigate drift from the first stage. We define a symmetric InfoNCE loss, given a minibatch of N matched pairs (q_i, k_i) (video–satellite or text–satellite), where $\ell(a, b)$ is the standard InfoNCE loss using cosine similarity and a temperature hyperparameter. We also define the regularization loss between the embeddings produced by Stage 1 and Stage 2 satellite encoders $\mathbf{f}_s^{\text{stage1}}$ and $\mathbf{f}_s^{\text{stage2}}$.

$$\mathcal{L}_{\text{ctr}} = \frac{1}{2N} \sum_{i=1}^{N} \left[\ell(q_i, k_i) + \ell(k_i, q_i)\right], \quad \mathcal{L}_{\text{reg}} = 1 - \text{sim}(\mathbf{f}_s^{\text{stage1}}, \mathbf{f}_s^{\text{stage2}}) \tag{5}$$

Finally, we freeze the TrajLoc encoders and train TrajMod's two MLPs using the same contrastive objective $\mathcal{L}_{\text{ctr}}$ for video–satellite and text–satellite alignment. Since both modalities are conditioned on the same $\mathbf{z}_{\text{traj}}$, adding a text–video contrastive term serves as a regularizer: it discourages modality-specific

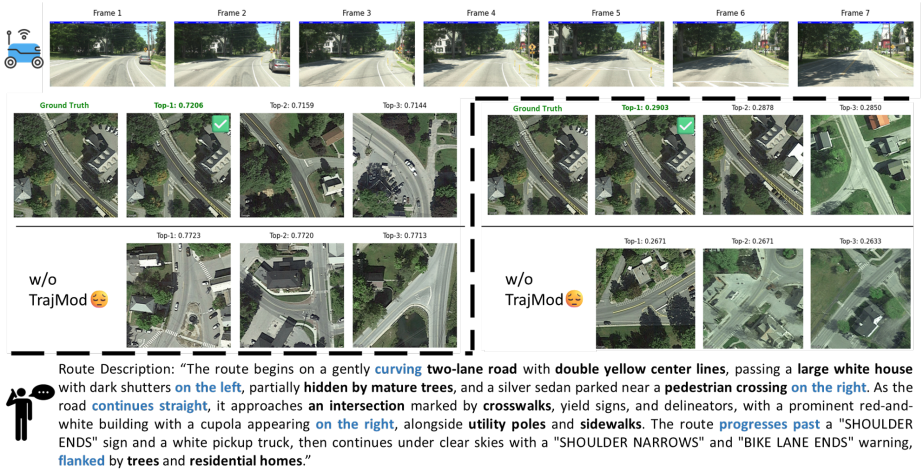

Fig. 5: Cross-view geo-localization with and without TrajMod. **Blue** highlights the spatial cues, while **bold** denotes the observed objects. As observed in Top-3 candidates, TrajMod enables more layout-aligned retrieval results.

noise that is not shared across views and encourages both $\mathbf{f}'_v$ and $\mathbf{f}'_t$ to preserve the common, geometry-grounded factors injected by $\mathbf{z}_{\text{traj}}$.

5 Experiments

We evaluate TrajLoc on the SeqGeo-VL dataset, assessing cross-view retrieval from both video and text queries. All experiments use Recall@K (R@K) as the primary metric and share a common training recipe, consistent except for a few noted exceptions, to ensure a fair comparison.

5.1 Experimental Setup and Implementation Details

We evaluate our framework on two tasks: cross-view text geo-localization [32], which retrieves a geo-referenced satellite image from a natural language route description, and cross-view video geo-localization [16,38], which matches ground-level videos to geo-referenced satellite imagery. We adopt the pretrained CLIP ViT-L/14 as the backbone for all three encoders. The model is optimized using Adam with an initial learning rate of 1×10^{-5}. Video- and text-alignment with satellite imagery are each trained for 40 epochs with a total batch size of 24, while TrajMod is trained for 20 epochs with a batch size of 128. Full implementation details of the compared models are provided in the supplementary material.

5.2 Cross-view Video Geo-localization Experiments

For cross-view video geo-localization, we compare our method with recent state-of-the-art (SOTA) models, including FlexGeo [25], GARet [16], and SeqGeo [38].

To mitigate the impact of backbone selection, we re-implemented SeqGeo using a CLIP ViT-L/14 architecture (denoted as SeqGeo†). Notably, our video encoding method employs simple mean pooling [13] for efficiency, whereas SeqGeo utilizes multiple transformer layers for feature aggregation.

As shown in Table 2, TrajLoc achieves superior performance on cross-view video geo-localization, reaching an R@1 of 12.09%. Upgrading the SeqGeo backbone from VGG16 to CLIP ViT-L/14 improves its R@1 from 1.80% to 8.14%; however, TrajLoc still outperforms it by a significant margin using the same ViT-L/14 backbone, even while operating with a lower computational budget.

We further evaluate Qwen3-VL-Embedding, as its MLLM-based pre-training on large-scale multimodal data provides strong video understanding. We apply LoRA fine-tuning to equip it with cross-view reasoning capabilities, since end-to-end adaptation of MLLMs is substantially more resource-intensive under practical constraints. However, it still underperforms our approach on this challenging task, as general-purpose VLM embeddings lack the spatially grounded characteristics essential for cross-view matching.

Table 2: Comparison of cross-view video geo-localization results. *denotes LoRA fine-tuning [8]; †denotes reimplementation using the original code. Otherwise, results are excerpted from the original papers.

Method	Backbone	R@1	R@5	R@10	R@1%
SeqGeo [38]	VGG16	1.80	6.45	10.36	34.38
SeqGeo† [38]	ViT-L/14	8.14	24.42	34.02	66.17
GARet [16]	DeiT-m	3.34	11.19	17.18	44.39
FlexGeo [25]	ConvNext-B	3.51	14.22	20.77	48.53
Qwen3-VL-Embedding* [2]	Qwen3-VL-2B	5.44	19.21	28.15	61.16
TrajLoc (Ours)	ViT-L/14	**12.09**	**35.77**	**47.68**	**80.21**

5.3 Cross-view Text Geo-localization Experiments

For cross-view text geo-localization, we compare our method with recent state-of-the-art (SOTA) models, including CrossText2Loc [32] and CLIP, EVA2-CLIP, and SigLIP as also analyzed by Ye et al. [32]. Since cross-view text geo-localization is a new task [32], limited prior work exists for comparison. We also include recent general multimodal encoders, Perception Encoder and Qwen3-VL-Embedding. Most baselines share a similar text encoder design of extending the context window by *linearly interpolating* positional embeddings.

Table 3 shows TrajLoc consistently achieves the best performance (R@1=2.52%, R@1%=45.48%), surpassing the strongest baselines by a large margin: 2.6× higher R@1 than CrossText2Loc (ViT-L/14@336) and 1.7× higher R@1% than EVA2-CLIP (ViT-L/14@336). Notably, scaling the backbone (B→L→SO400M)

or increasing input resolution (@336/@384) yields only modest gains for CLIP-style methods, suggesting that the bottleneck lies not in backbone capacity but in the spatial reasoning required to bridge egocentric route descriptions and allocentric satellite layouts. Overall, the results indicate that current VLM/MLLM embeddings still struggle with egocentric-to-allocentric spatial reasoning, motivating TrajLoc's trajectory-conditioned design to explicitly bridge narrative semantics with route geometry for more precise cross-view retrieval.

Table 3: Comparison of cross-view text geo-localization results. *denotes LoRA fine-tuning [8]; otherwise, all models are fully fine-tuned on SeqGeo-VL.

Method	Backbone	R@1	R@5	R@10	R@1%
CLIP [17]	ViT-B/16	0.53	1.58	3.33	15.94
CLIP [17]	ViT-L/14	0.80	3.32	6.01	22.99
EVA2-CLIP [6]	ViT-B/16	0.63	2.20	3.98	17.59
EVA2-CLIP [6]	ViT-L/14@336	0.89	3.78	6.70	26.99
SigLIP [36]	ViT-B/16	0.54	2.21	4.07	18.22
SigLIP [36]	ViT-L/16@384	0.85	3.13	5.24	21.33
SigLIP [36]	ViT-SO400M/14	0.80	2.92	5.24	21.86
Perception Enc. [3]	ViT-L/14@336	0.66	2.08	3.80	15.17
Qwen3-VL-Embed.* [2]	Qwen3-VL-2B	0.93	3.38	5.55	20.91
CrossText2Loc [32]	ViT-L/14	0.84	3.29	5.73	22.29
CrossText2Loc [32]	ViT-L/14@336	0.98	4.08	7.08	25.45
TrajLoc (Ours)	ViT-L/14	**2.52**	**9.82**	**16.11**	**45.48**

5.4 Ablation study

Impact of Temporal Length and Trajectory Geometry. To investigate the interplay between sequence length and trajectory conditioning, we evaluate video-to-satellite retrieval using varying numbers of input frames, with and without TrajMod. As shown in Fig. 6, performance generally improves across all metrics (R@1, R@5, and R@10) as the number of frames grows from 1 to 6, validating our core motivation that extended temporal context provides richer cues for cross-view geo-localization compared to single-frame observations. Moreover, the trajectory-conditioned variant (*w/ TrajMod*) consistently outperforms the baseline (*w/o TrajMod*), with the performance gap steadily widening as sequences lengthen—e.g., the R@1 margin (Δ) grows from 2.6% at 1 frame to 5.6% at 6 frames. We attribute this to the fact that longer sequential observations inevitably suffer from stronger visual aliasing and feature smoothing. Rather than relying on temporal modeling or heuristic frame selection, TrajMod modulates semantic features conditioned on the geometry of the viewpoint trajectory, capturing the spatial layout of observed semantics. The qualitative results are shown in Fig. 5.

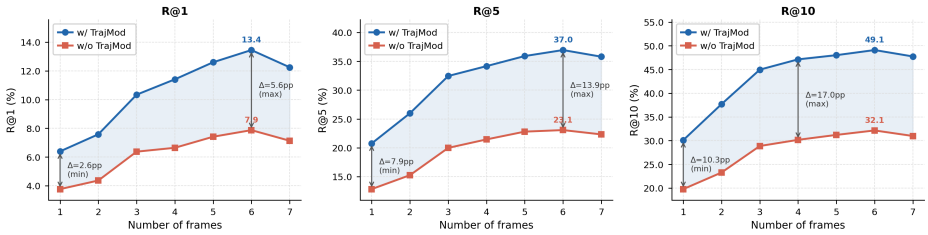

Fig. 6: Video geo-localization performance with varying frame number and TrajMod.

Location Prior Context. In practical applications, users or mobile robots often operate with a location prior, such as knowing their rough geographic district when GPS signals are severely obscured. To simulate these conditions, we evaluate retrieval performance under varying gallery sizes, effectively narrowing the search space based on prior knowledge. As shown in Fig. 7, restricting the candidate pool substantially improves the absolute retrieval accuracy and demonstrates the practical utility of our explicitly grounded embeddings.

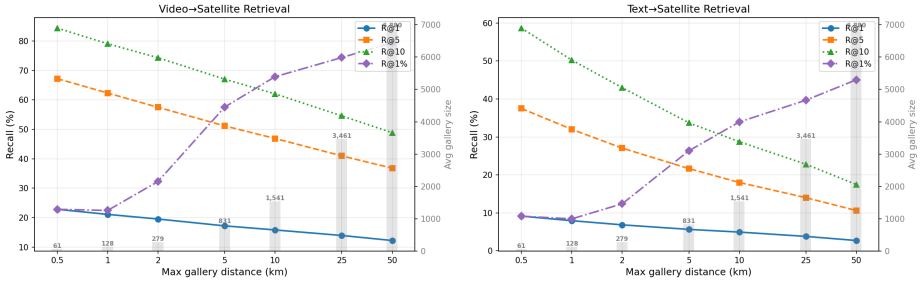

Fig. 7: Performance with different gallery sizes given a location prior.

Synergy between Co-training and TrajMod. We investigate the mutual benefits of jointly training cross-view geo-localization with visual or linguistic sequential observations, and how TrajMod amplifies this synergy. As reported in Table 4, without TrajMod, simply co-training the two modalities yields only marginal changes over independent training (e.g., Video R@1 improves slightly from 6.87% to 7.27%, and Text R@1 from 0.80% to 0.98%).

However, when TrajMod is introduced, the gains from co-training are significantly amplified. For the video-to-satellite task, co-training with the text modality boosts R@1 by a remarkable 2.40% (from 9.69% to 12.09%). Similarly, for text-to-satellite, co-training with video improves R@1 from 1.90% to 2.52%. We attribute this synergy to TrajMod's explicit geometric grounding. By anchoring visual observations and textual narratives to the same trajectory geometry, TrajMod establishes a shared spatial representation that facilitates cross-modal

knowledge transfer, allowing visual details and route-level semantics to mutually reinforce each other.

Table 4: Ablation study on co-training components and TrajMod.

	Video Geo-localization				Text Geo-localization			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
(a) TrajMod component ablation								
Full model	**12.09**	**35.77**	**47.68**	**80.21**	**2.52**	**9.82**	**16.11**	45.48
w/o Txt2Vid contrastive loss	10.23	30.57	41.21	74.20	2.17	8.49	14.19	42.22
w/o TrajMod	7.27	22.13	30.92	63.47	0.98	4.16	7.21	26.15
(b) Co-training ablation (w/ TrajMod)								
Full model	**12.09**	**35.77**	**47.68**	**80.21**	**2.52**	**9.82**	**16.11**	45.48
Vid2Sat only	9.69	29.75	40.30	73.60	–	–	–	–
Txt2Sat only	–	–	–	–	1.90	9.12	15.57	**47.25**
(c) Co-training ablation (w/o TrajMod)								
Full model	7.27	22.13	30.92	63.47	0.98	4.16	7.21	26.15
Vid2Sat only	6.87	22.50	31.60	64.09	–	–	–	–
Txt2Sat only	–	–	–	–	0.80	3.32	6.01	22.99
w/o Curriculum learning	5.01	18.28	27.25	61.81	0.86	3.86	6.52	24.40

Utilizing Trajectory Geometry. We compare two methods for incorporating the same trajectory information: text prompting for MLLM-based embeddings (e.g., Qwen3-VL) and TrajMod for our CLIP-style co-trained architectures (Table 6). While adding trajectory details via prompts yields a modest improvement (+1.34% for R@1%), TrajMod provides a substantial gain of +19.33% for R@1%. This demonstrates that explicit geometric modulation is significantly more effective than text-based instructions for spatial reasoning in cross-view retrieval.

Co-sequenced Multimodal Query. We further examine the performance of our unified approach when both video and text queries are available. Linearly blending video and text query embeddings ($\alpha = 0.4$) consistently outperforms the stronger single-modality baseline across all frame counts, with R@10 gains ranging from +9.49 at 1 frame down to +4.75 at 7 frames; the complete per-frame results are reported in the supplementary material. This indicates text and video provide complementary route cues, with text especially valuable when video is sparse, missing, or corrupted—a favorable storage/encoding trade-off.

5.5 Discussion on SeqGeo-VL

To evaluate the necessity of sequential observations and spatiotemporal cues, we compare SeqGeo-VL against two variants (Table 5) with CrossText2Loc [32].

First, *Single-view Description* uses only the VLM-generated description of the final waypoint. This yields the lowest performance (10.63% R@1%), indicating a single-frame observation lacks the context for unambiguous matching.

Second, to verify the importance of temporal order, we design *Sequential-view Objects*. To construct this variant, we extract perceived entities and nouns from

the route descriptions and order them chronologically to maintain their temporal order of appearance. Despite its minimal length (26.3 words), it significantly outperforms the single-view baseline (17.41% vs. 10.63% R@1%), validating that sequential observations provide critical localization cues.

Finally, the full SeqGeo-VL achieves the best results (22.29% R@1%). The performance gap between the object-only variant and the full dataset demonstrates that relational context (i.e., spatiotemporal cues) is essential for constructing spatially-aware language representations. SeqGeo-VL also exhibits a lower average similarity, indicating more diverse textual descriptions.

Table 5: Text quality and retrieval results.

Dataset	Text quality		Retrieval	
	Word Len.	Avg Sim.	R@1	R@1%
Single-view caption	50.9	0.270	0.36	10.63
Sequential-view objects	26.3	0.281	0.58	17.41
SeqGeo-VL	**101.3**	**0.247**	**0.84**	**22.29**

Table 6: Comparison of trajectory conditioning mechanisms.

Method	R@1%	Δ
Qwen3-VL-Embed. [2]	20.91	-
+ Prompt	22.25	+1.34
Our CLIP-style baseline	26.15	-
+ TrajMod	**45.48**	**+19.33**

6 Conclusion

We extend route-level cross-view geo-localization beyond video queries to natural-language route descriptions. We introduce SeqGeo-VL, a multimodal benchmark of video-text-satellite triplets for this task; TrajLoc, a unified approach for video and text queries; and TrajMod, a lightweight module that incorporates explicit trajectory geometry for feature modulation. TrajLoc achieves state-of-the-art performance on cross-view geo-localization from both videos and route descriptions. These results demonstrate the value of jointly modeling sequential visual and linguistic observations, and highlight the role of trajectory geometry in complementing vision-language representations for cross-view spatial reasoning.

Acknowledgements

This research used the TGI RAILS advanced compute and data resource which is supported by the National Science Foundation (award OAC-2232860) and the Taylor Geospatial Institute.

References

1. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3674–3683 (2018)

2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. arXiv preprint arXiv:2504.13181 (2025)
4. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. Advances in Neural Information Processing Systems **37**, 135062–135093 (2024)
5. Chu, M., Zheng, Z., Ji, W., Wang, T., Chua, T.S.: Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In: European Conference on Computer Vision. pp. 213–231. Springer (2024)
6. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. Image and Vision Computing **149**, 105171 (2024)
7. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
9. Kolmet, M., Zhou, Q., Ošep, A., Leal-Taixé, L.: Text2pos: Text-to-point-cloud cross-modal localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6687–6696 (2022)
10. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 891–898 (2013)
11. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geolocalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5624–5633 (2019)
12. Liu, M., He, H., Ricci, E., Wu, W., Zhou, B.: Urbanverse: Scaling urban simulation by watching city-tour videos. In: The Fourteenth International Conference on Learning Representations (2026)
13. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: CLIP4Clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)
14. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
15. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
16. Pillai, M.S., Rizve, M.N., Shah, M.: Garet: cross-view video geolocalization with adapters and auto-regressive transformers. In: European Conference on Computer Vision. pp. 466–483. Springer (2024)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
18. Sastry, S., Choudhury, A., Madala, P., Su, G.M.: Probabilistic multimodal learning with bayesian disagreements. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 7546–7555 (June 2026)

19. Shi, Y., Yu, X., Wang, S., Li, H.: Cvlnet: Cross-view semantic correspondence learning for video-based camera localization. In: Asian Conference on Computer Vision. pp. 123–141. Springer (2022)
20. Tian, H., Meng, J., Zheng, W.S., Li, Y.M., Yan, J., Zhang, Y.: Loc4plan: Locating before planning for outdoor vision and language navigation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 4073–4081 (2024)
21. Vyas, S., Chen, C., Shah, M.: Gama: Cross-view video geo-localization. In: European Conference on Computer Vision. pp. 440–456. Springer (2022)
22. Wang, Z., Zhou, S., He, S., Huang, H., Yang, L., Zhang, Z., Cheng, X., Ji, S., Jin, T., Zhao, H., et al.: Spatialclip: Learning 3d-aware image representations from spatially discriminative language. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29656–29666 (2025)
23. Wei, W., Gui, Z., Peng, D., Ye, T., Wu, H.: Variational adapter for cross-modal similarity representation. In: Forty-third International Conference on Machine Learning (2026), https://openreview.net/forum?id=BkZ40FNKlL
24. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3961–3969 (2015)
25. Wu, Q., Xia, P., Yu, L., Liu, Y., Xiong, M., Zhong, L., Chen, J., Yang, M., Zhang, Y., Wan, Y.: Set-cvgl: A new perspective on cross-view geo-localization with unordered ground-view image sets. ISPRS Journal of Photogrammetry and Remote Sensing **233**, 328–345 (2026). https://doi.org/https://doi.org/10.1016/j.isprsjprs.2026.01.037, https://www.sciencedirect.com/science/article/pii/S0924271626000456
26. Xia, P., Yu, L., Wan, Y., Wu, Q., Chen, P., Zhong, L., Yao, Y., Wei, D., Liu, X., Ru, L., et al.: Cross-view geo-localization with panoramic street-view and vhr satellite imagery in decentrality settings. ISPRS Journal of Photogrammetry and Remote Sensing **227**, 1–11 (2025)
27. Xia, Y., Shi, L., Ding, Z., Henriques, J.F., Cremers, D.: Text2loc: 3d point cloud localization from natural language. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14958–14967 (2024)
28. Xing, E., Stylianou, A., Pless, R., Jacobs, N.: Quari: Query adaptive retrieval improvement. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 62890–62911. Curran Associates, Inc. (2025), https://proceedings.neurips.cc/paper_files/paper/2025/file/5ae6634cb92414dc34d24e863e8d0809-Paper-Conference.pdf
29. Xiong, Z., Xing, X., Workman, S., Khanal, S., Jacobs, N.: Mixed-view panorama synthesis using geospatially guided diffusion. Transactions on Machine Learning Research
30. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
31. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
32. Ye, J., Lin, H., Ou, L., Chen, D., Wang, Z., Zhu, Q., He, C., Li, W.: Where am i? cross-view geo-localization with natural language descriptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5890–5900 (2025)

33. Ye, J., Lv, Z., Li, W., Yu, J., Yang, H., Zhong, H., He, C.: Cross-view image geo-localization with panorama-bev co-retrieval network. In: European Conference on Computer Vision. pp. 74–90. Springer (2024)
34. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)
35. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? arXiv preprint arXiv:2210.01936 (2022)
36. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
37. Zhang, J., Yu, X., Fang, Y., Stouffs, R., Trivic, Z.: Urbangraphembeddings: Learning and evaluating spatially grounded multimodal embeddings for urban science. arXiv preprint arXiv:2602.08342 (2026)
38. Zhang, X., Sultani, W., Wshah, S.: Cross-view image sequence geo-localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2914–2923 (2023)
39. Zheng, Z., Wei, Y., Yang, Y.: University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: Proceedings of the 28th ACM international conference on Multimedia. pp. 1395–1403 (2020)
40. Zhu, S., Shah, M., Chen, C.: Transgeo: Transformer is all you need for cross-view image geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1162–1171 (2022)
41. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3640–3649 (2021)