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

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

A Overview

This appendix provides additional details and analyses that complement the main paper.
Dataset Construction (Sec. B) describes the proposed annotation pipeline, including the VLM and LLM prompts used for semantic extraction and route summarization, and reports the statistics of SeqGeo-VL.
Implementation Details (Sec. C) gives the full training recipe for TrajLoc and the hyperparameters of all compared baselines.
Additional Ablations (Sec. D) presents further analyses on:

- Regularization loss for the satellite encoder (Sec. D.1).
- More design choices for TrajMod and its stability analysis (Sec. D.2).
- Robustness under adverse weather conditions (Sec. D.3).
- Robustness under incomplete route descriptions (Sec. D.4).
- Robustness under heading perturbation (Sec. D.5).
- Co-sequenced Multimodal Query (Sec. D.6).

Inference Efficiency (Sec. E) evaluates query-encoding throughput.
Qualitative Results (Sec. F) shows dataset samples, successful examples for both video and text geo-localization, and a representative failure case.

B Dataset Construction

To ensure high-quality trajectory annotations, we use a progressive pipeline: a frame-level VLM for visual semantic extraction and a trajectory-level LLM for route summarization. This design lets the VLM focus on visual recognition while the LLM handles higher-level reasoning, and it also allows us to inspect intermediate captions and refine prompts at each stage.

B.1 Data Annotation Prompts

In this section, we present the detailed prompts used in our annotation pipeline.

Motion Context Generation. We derive motion context from the trajectory heading metadata (h_i) and incorporate it into prompts during the Semantic Extraction and Structural Summarization stages, respectively.

(1) *{turn_hint}* represents the frame-wise relative change: the first frame is labeled as *starting*, while subsequent frames are labeled as *straight* ($|\Delta h| < \tau$), *left* ($\Delta h \leq -\tau$), or *right* ($\Delta h \geq \tau$) based on the heading difference from the previous frame. $\tau = 30°$.

(2) *{overall_turn_info}* provides a global summary by comparing the final and initial headings, describing the overall maneuver and total heading change. All heading differences are normalized to $[-180°, 180°]$.

VLM Prompt for Semantic Extraction (Per Frame)

Instruction:
Describe what you see in this street-view image in one to two concise sentences.
Turn context: *{turn_hint}*
Focus on: road features (lanes, markings, surface), intersections, traffic signs, buildings, landmarks, vegetation, or other distinctive visual elements.
Constraint: Be specific and objective. Avoid generic descriptions and weather/lighting conditions.

The VLM output, `{image_descriptions}`, contains concatenated descriptions of all sequential frames and is passed to the next stage, where the LLM produces a coherent summary.

LLM Prompt for Structural Summarization

Instruction:
You are a helpful assistant that summarizes trajectory route descriptions.

Below are sequential street-view image descriptions from a driving trajectory. Summarize this route in ONLY TWO TO THREE sentences, capturing the key visual features and route progression in order.
FOCUS ON: distinctive landmarks, road features, intersections, and overall route character.
DO NOT FOCUS ON: weather, sky colors, lighting conditions, dates or generic elements.

Overall trajectory: *{overall_turn_info}*
{image_descriptions}

Provide a coherent, flowing summary that someone could use to recognize this route.

B.2 Dataset Statistics

Figure S1 provides a broader view of SeqGeo-VL. First, the train and validation sets are drawn from the same broad region, but their sampled trajectories

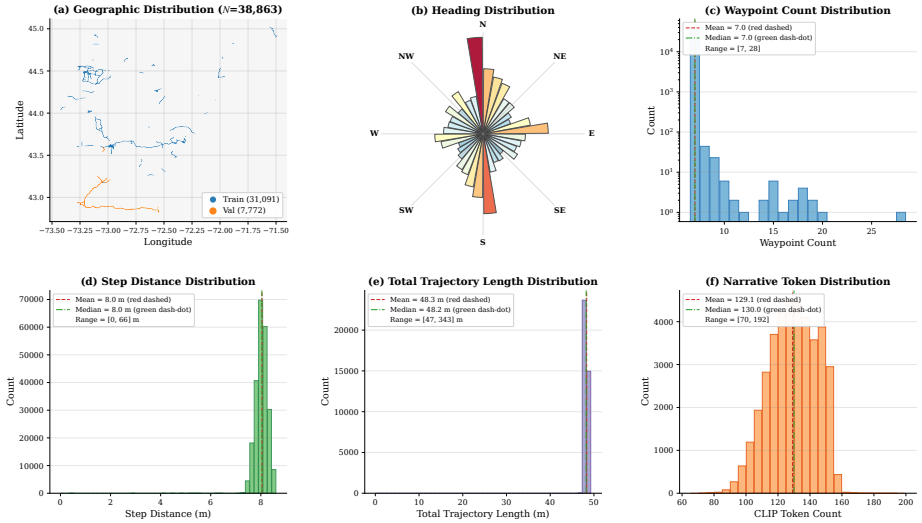

Fig. S1: Comprehensive statistics of SeqGeo-VL: (a) geographic distribution of trajectory centers in train/val splits, (b) heading distribution, (c) waypoint-count distribution, (d) step-distance distribution, (e) total trajectory-length distribution, and (f) narrative token-count distribution.

are spatially disjoint (no direct overlap). This split design preserves a consistent large-scale geographic context while preventing leakage from near-duplicate locations across splits. Second, the heading histogram is approximately bi-modal with dominant North/South directions, which is consistent with road-network anisotropy. Third, trajectory geometry is relatively compact: most samples contain 7 waypoints (range: 7–28), with short per-step displacement (mean: 8.0 m) and total path length concentrated around 48.3 m. Finally, route descriptions are substantially longer than vanilla CLIP text capacity (mean: 129.1 tokens, median: 130 tokens, range: 70–192 tokens), supporting our choice to expand text positional embeddings and to model long-form sequential semantics.

Three trained annotators scored a geographically uniform ∼1% subset (400 trajectories) on object fidelity (OF), spatiotemporal consistency (ST), and route completeness (RC), with inter-annotator agreement measured by Gwet's AC2: OF 2.77/3 (AC2= 0.89), ST 2.72/3 (AC2= 0.90). A sample passes iff RC is rated complete by a majority and $\mathrm{OF} + \mathrm{ST} > 3$, giving a pass rate of 91.8%.

C Implementation Details

Here we further explain the implementation details of our approach and compared ones.

Three-stage Training. As discussed in Sec. 4.3 of the main paper, TrajLoc is trained with the following objectives:

$$\mathcal{L}^{(1)} = \mathcal{L}_{\text{ctr}}(\mathbf{f}_v, \mathbf{f}_s^{(1)}) \tag{1}$$

$$\mathcal{L}^{(2)} = \mathcal{L}_{\text{ctr}}(\mathbf{f}_t, \mathbf{f}_s^{(2)}) + \alpha_{\text{reg}}\mathcal{L}_{\text{reg}}(\mathbf{f}_s^{(1)}, \mathbf{f}_s^{(2)}) \tag{2}$$

$$\mathcal{L}^{(3)} = \mathcal{L}_{\text{ctr}}(\mathbf{f}'_v, \mathbf{f}_s^{(2)}) + \mathcal{L}_{\text{ctr}}(\mathbf{f}'_t, \mathbf{f}_s^{(2)}) + \alpha_{\text{align}}\mathcal{L}_{\text{ctr}}(\mathbf{f}'_v, \mathbf{f}'_t) \tag{3}$$

where $\mathbf{f}_v{=}E_v(v)$, $\mathbf{f}_t{=}E_t(t)$ denote video and text features, and $\mathbf{f}_s^{(k)}{=}E_s^{(k)}(s)$ denotes satellite features at stage k. Stages (1) and (2) follow a curriculum that gradually increases alignment difficulty, while Stage (3) freezes all encoders and trains only the TrajMod module, where $\mathbf{f}'_v$ and $\mathbf{f}'_t$ are TrajMod-conditioned features. The text–video contrastive term in $\mathcal{L}^{(3)}$, weighted by α_{align}, regularizes TrajMod to preserve the shared geometry-grounded factors injected by the trajectory geometry embedding. We set $\alpha_{\text{reg}}{=}0.5$ and $\alpha_{\text{align}}{=}0.1$.

All three encoders are initialized from CLIP ViT-L/14 [2] with text positional embeddings interpolated from 77 to 150 tokens following [4]. Both street-view and satellite images are resized to 224×224 using the default CLIP preprocessing.

We use the Adam optimizer with a cosine annealing schedule and mixed-precision training throughout. Stages (1) and (2) each train for 40 epochs with learning rate=$1{\times}10^{-5}$ and batch size 24. Stage (3) trains TrajMod for 20 epochs with batch size 128. Each trajectory is represented by 7 street-view frames, uniformly sampled at test time and randomly sampled during training.

Compared Approaches. Unless otherwise noted, all reimplemented baselines share the same training hyperparameters with TrajLoc: Adam optimizer, learning rate=$1{\times}10^{-5}$, batch size=24, 40 epochs, cosine annealing schedule, and mixed-precision training on a single H100/A100 GPU. For a fair comparison on text geo-localization, all CLIP-based baselines also apply positional embedding interpolation to accommodate longer route descriptions (Qwen3-VL-Embedding does not require this, as its LLM backbone already supports a sufficiently large context window). Below we highlight only the differences specific to each method.

CrossText2Loc [4]. We reimplement CrossText2Loc using CLIP ViT-L/14@336 and ViT-L/14 backbones with positional embeddings expanded to 300 tokens, following the original paper. The model is trained with a text-to-satellite symmetric contrastive loss for 40 epochs (learning rate=$1{\times}10^{-5}$, batch size=128, following the original implementation).

SeqGeo† [5]. We reimplement SeqGeo by replacing the original VGG16 backbone with CLIP ViT-L/14 for both the street-view and satellite feature extractors, while retaining the original temporal aggregation design: a 6-layer Transformer encoder (8 heads, dropout=0.3) with sinusoidal positional encoding and random frame masking (up to 6 of 7 frames) during training, followed by average pooling. The model is trained with the soft-margin triplet loss [5] (γ=10) for 50 epochs using Adam (learning rate=$1{\times}10^{-5}$, batch size=24) with linear learning rate decay starting at epoch 30, following the original implementation.

Qwen3-VL-Embed. [1] We fine-tune the Qwen3-VL-Embedding-2B with LoRA applied to the Q/K/V projections and all MLP layers. We set the rank to 32,

α to 32, and dropout to 0.05. Embeddings are extracted via last-token pooling with L2 normalization. We use modality-specific instructions (*e.g.*, *"Represent this top-down satellite image."* for satellites, *"Represent this streetview video."* for videos, and *"Represent this trajectory description."* for route descriptions). For the trajectory-aware variant (Table 6 in the main paper), the instruction includes angle information (*e.g.*, *"...with an overall bearing of 92.3 degrees. Headings per frame: [91.0, 92.5, ...]"*). The model is trained for 40 epochs with learning rate=1×10^{-4}, batch size=8 for video geo-localization due to memory constraints, and 40 epochs with learning rate=1×10^{-5}, batch size=24 for text-based geo-localization.

D Additional Details of TrajLoc

Here we provide additional details of the TrajLoc model, including the role of the regularization loss, the representation of trajectory geometry, and the robustness of TrajLoc under different weather conditions, incomplete route descriptions, and heading perturbation.

D.1 Role of Regularization Loss

Table S1 studies how the Stage 1 satellite encoder should be handled in Stage 2. Freezing it preserves the video-aligned retrieval space but limits text adaptation, while unconstrained fine-tuning causes drift and harms video geo-localization. The reported numbers are based on TrajLoc without TrajMod. Regularized tuning provides the best trade-off, retaining the Stage 1 structure while adapting to text supervision.

This experiment shows that the regularization loss prevents catastrophic drift of the satellite encoder during Stage 2, enabling text geo-localization without sacrificing video performance.

Table S1: Role of regularization loss. Ablation on the Stage 2 satellite encoder configuration: frozen, tunable, and tunable with regularization. Regularized tuning best preserves video geo-localization while improving text geo-localization.

	Video Geo-localization				**Text Geo-localization**			
Method	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
Frozen Satellite Encoder	6.87	**22.50**	**31.60**	**64.09**	0.87	3.05	5.31	21.38
Tunable Satellite Encoder	4.21	13.83	20.54	51.40	**1.07**	3.58	6.14	23.33
Tunable w/ Regularization	**7.27**	22.13	30.92	63.47	0.98	**4.16**	**7.21**	**26.15**

D.2 Ablation Studies and Stability Analysis of TrajMod

Table S2 and Table S3 validate our choices of trajectory geometry representations. As shown in Table S2, replacing raw angles with Fourier feature mapping

yields consistent gains across all metrics for both video and text geo-localization (e.g., Video R@1: 7.80→12.09, Text R@1: 1.17→2.52). This suggests that directly using raw angles is less effective for optimization, while Fourier features provide a more expressive representation for trajectory cues and improve the modulation of vision-language embeddings.

Table S2: Ablation study on representation of trajectory geometry

	Video Geo-localization				**Text Geo-localization**			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
Raw angles	7.80	23.60	32.37	64.85	1.17	4.71	7.84	27.90
Fourier feature mapping	**12.09**	**35.77**	**47.68**	**80.21**	**2.52**	**9.82**	**16.11**	**45.48**

Headings and OD bearing carry largely overlapping information, since the OD direction can be recovered from the aggregate trend of waypoint headings. As shown in Table S3, either cue alone already yields strong performance, with full headings being slightly more informative on a turn-rich subset (first heading and OD bearing differing by $> 10°$) where local turning patterns matter. We keep both headings and OD bearing for robustness: under $\pm 30°$ test-time noise, the joint variant degrades far more gracefully, as one corrupted cue still leaves the other as a stable fallback.

Table S3: Fine-grained ablations of trajectory geometry components (R@10 for video geo-localization).

Mode	Clean	+ Head. noise	+ Bear. noise	Turn-rich
Both	<u>47.89</u>	**31.23**	**45.82**	**58.52**
Headings	**48.02**	27.73	–	<u>57.95</u>
Bearings	46.85	–	27.59	55.68

We re-trained trajectory-conditioned modulation with five different seeds (Table S4); small standard deviations show the gains are stable.

Table S4: Performance stability across five runs (mean ± std.).

Query	R@1	R@5	R@10	R@1%
Video	$12.37_{\pm 0.10}$	$35.90_{\pm 0.37}$	$48.14_{\pm 0.25}$	$80.36_{\pm 0.22}$
Text	$2.51_{\pm 0.05}$	$9.53_{\pm 0.14}$	$15.95_{\pm 0.28}$	$45.24_{\pm 0.54}$

Table S5: Robustness under adverse weather conditions. We report R@1 and R@1% for the baseline and **TrajMod** under various environmental perturbations. Numbers in gray and blue indicate **retention rates** relative to the clean setting. R@5 and R@10 follow the same trend and are omitted for readability.

	R@1 ↑		R@1% ↑	
Condition	Baseline	+ TrajMod	Baseline	+ TrajMod
Clean	7.13	**12.25**	63.54	**80.20**
Dark	5.25 (73.6)	**9.93** (**81.1**)	59.78 (94.1)	**78.80** (**98.3**)
Fog	3.95 (55.4)	**7.51** (**61.3**)	53.41 (84.1)	**70.66** (**88.1**)
Rain	5.11 (71.7)	**8.80** (**71.8**)	56.48 (88.9)	**75.27** (**93.9**)
Snow	6.46 (90.6)	**11.13** (**90.9**)	61.90 (97.4)	**80.48** (**100.3**)
Low Light	2.42 (33.9)	**4.98** (**40.7**)	39.01 (61.4)	**59.93** (**74.7**)
Fog + Rain	3.28 (46.0)	**6.09** (**49.7**)	47.23 (74.3)	**65.07** (**81.1**)
Fog + Snow	2.56 (35.9)	**4.76** (**38.9**)	41.41 (65.2)	**60.50** (**75.4**)
Rain + Snow	5.35 (75.0)	**9.60** (**78.4**)	59.15 (93.1)	**77.38** (**96.5**)

D.3 Robustness of TrajLoc under Visually-degraded Conditions

We evaluate the robustness of TrajLoc under adverse weather conditions following the setup in [3]. Table S5 reports both absolute scores and the *performance retention rate* (the ratio of the perturbed metric to its clean counterpart) for each setting.

Across all weather conditions—single perturbations (Dark, Fog, Rain, Snow, Low Light) and compound perturbations (Fog+Rain, Fog+Snow, Rain+Snow)—TrajMod consistently achieves higher retention rates than the baseline model (w/o TrajMod), demonstrating that injecting trajectory geometry improves robustness to visual degradation. The advantage is most pronounced under severe conditions that heavily corrupt visual appearance: under Low Light, the baseline retains only 33.9% of its clean R@1 while TrajMod retains 40.7%; for R@1%, the gap widens to 61.4% vs. 74.7% (+13.3 pp). Similarly, under Fog+Snow, TrajMod preserves 75.4% of R@1% compared to 65.2% for the baseline (+10.2 pp). Even under mild perturbations such as Snow, TrajMod achieves near-perfect retention (R@1%: 100.3%), indicating full compensation.

This behavior is consistent with the expectation that weather conditions primarily degrade pixel-level appearance while leaving trajectory geometry—heading angles, bearing directions, and spatial layout—largely unaffected. The consistent retention-rate advantage across all weather types demonstrates that trajectory geometry provides a practically useful complementary signal for cross-view retrieval under real-world visual degradation.

D.4 Robustness of TrajLoc under Incomplete Route Descriptions

We evaluate TrajLoc's robustness when the textual route description is incomplete by randomly dropping tokens from the input text with probability p. Ta-

ble S6 compares the baseline (without TrajMod) and the full model with TrajMod under increasing dropout rates.

Even with 30% of tokens removed, TrajMod retains 67.6% of its clean R@1 compared to only 53.0% for the baseline, and 80.8% of R@1% versus 75.2%. This experiment simulates the scenario where the text geo-localization may be brittle to noisy or incomplete user inputs. TrajMod's geometric modulation provides an information channel that is orthogonal to lexical content, improving graceful degradation under text corruption.

Table S6: Robustness under incomplete route descriptions. Text-to-satellite geo-localization under random token dropout with varying dropout probability p. Numbers in gray and blue indicate **retention rates** relative to the clean setting. R@5 and R@10 follow the same trend and are omitted for readability.

	R@1 ↑		**R@1%** ↑	
Dropout p	Baseline	+ TrajMod	Baseline	+ TrajMod
0 (Clean)	1.00	**2.53**	26.15	**45.50**
0.05	0.73 (73.0)	**2.29** (**90.5**)	24.81 (94.9)	**43.75** (**96.2**)
0.10	0.78 (78.0)	**2.03** (**80.2**)	24.06 (92.0)	**42.86** (**94.2**)
0.20	0.66 (66.0)	**1.79** (**70.8**)	21.42 (81.9)	**39.14** (**86.0**)
0.30	0.53 (53.0)	**1.71** (**67.6**)	19.66 (75.2)	**36.75** (**80.8**)

D.5 Robustness of TrajLoc under Heading Perturbation

TrajMod relies on trajectory geometry, specifically waypoint headings and the OD bearing, to modulate the query embeddings of sequential observations. In practice, these signals may be corrupted by sensor noise or localization drift at inference time. We evaluate TrajLoc's sensitivity to such noise under three perturbation modes: **Bearing Only** (perturbing the global bearing between consecutive waypoints, simulating positional drift), **Heading Only** (perturbing the camera heading at each observation, simulating compass noise), and **Bearing + Heading** (perturbing both simultaneously). For each mode, we apply fixed angular offsets of $+\delta$ and $-\delta$ and report the average performance over the two directions. Table S7 reports R@1 for video-to-satellite and text-to-satellite geo-localization along with performance retention rates.

Bearing perturbation has minimal impact: even at δ=30°, V→S R@1 retains 93.7% and T→S R@1 retains 95.3% of clean performance. In contrast, heading perturbation causes significant degradation—V→S R@1 drops to 56.4% retention at δ=30°. This asymmetry is primarily due to differences in information density: heading is a per-frame signal that provides dense orientation supervision, whereas bearing is computed only from the start and end points of the trajectory and is therefore much sparser. The model naturally learns to rely

Table S7: Robustness under heading perturbation. R@1 (%) under three orientation perturbation modes with varying delta angle δ. Numbers in gray indicate **retention rates** relative to δ=0°.

	Bearing Only		**Heading Only**		**Bearing + Heading**	
δ	V→S	T→S	V→S	T→S	V→S	T→S
0° (Clean)	**12.25**	**2.53**	**12.25**	**2.53**	**12.25**	**2.53**
±7.5°	12.11 (98.9)	2.55 (100.8)	11.32 (92.4)	2.30 (90.9)	11.08 (90.4)	2.19 (86.6)
±15°	11.82 (96.5)	2.53 (100.0)	9.68 (79.0)	1.80 (71.1)	8.91 (72.7)	1.59 (62.8)
±22.5°	11.59 (94.6)	2.49 (98.4)	7.91 (64.6)	1.33 (52.6)	7.06 (57.6)	1.15 (45.5)
±30°	11.48 (93.7)	2.41 (95.3)	6.91 (56.4)	0.95 (37.5)	5.84 (47.7)	0.72 (28.5)

more heavily on the denser heading cue during training, making it sensitive to heading noise but resilient to bearing noise.

D.6 Co-sequenced Multimodal Query

As shown in Table S8, we 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.

Table S8: R@10 between video-only, text-only, and video-text query blending.

#Frames	Video-only	Text-only	Video-text blend	Δ
1	30.11	15.08	**39.60**	+9.49
3	44.98	15.97	**50.31**	+5.33
5	48.02	15.99	**52.79**	+4.77
7	47.75	16.06	**52.50**	+4.75

E Inference Efficiency

We report the query-encoding throughput of different methods in Table S9. TrajLoc retains 92%/89% of the video/text encoding throughput of CLIP-L/14, indicating that its performance gains do not come at the cost of expensive inference. The modest overhead is mainly introduced by the lightweight TrajMod.

F Qualitative Results

We present dataset samples, retrieval examples for cross-view video and text geo-localization, and a failure case analysis.

Table S9: Query-encoding throughput (samples/s). Gray values denote relative speed compared to CLIP-L/14.

Encoder	Video query	Text query
Qwen3-VL-Emb.	6.36 (0.05×)	45.06 (0.17×)
SeqGeo (ViT-L/14)	101.73 (0.78×)	–
CrossText2Loc	–	251.71 (0.95×)
CLIP-L/14	130.64 (1.00×)	265.18 (1.00×)
Ours	119.74 (0.92×)	237.13 (0.89×)

F.1 Dataset Samples

Figure S2 shows representative samples from the dataset under complex environments. Each sample combines a satellite view, multiple sequential street-view observations, and route text; the multi-observation cues provide richer scene evidence, while their sequential relationships imply a spatial distribution prior over the underlying trajectory.

F.2 Successful Localization

Video geo-localization. Figure S3 presents a case where TrajMod enables correct Rank 1 localization. The top row shows the input ground-level video frames with the trajectory overlay. With TrajMod, the ground-truth satellite image is matched at Rank 1, whereas the baseline without TrajMod ranks it at Rank 8. The trajectory geometry helps disambiguate visually similar satellite tiles by encoding the spatial layout and heading direction of the route.

Text geo-localization. Figure S4 shows a successful text-to-satellite localization. The input is a natural language description of a driving route that mentions specific landmarks (e.g., a Shell gas station, Cumberland Farms, a Redbox kiosk) along with spatial cues such as road curvature and heading direction. With TrajMod, the ground-truth satellite tile is localized at Rank 1 with the highest similarity score, while the baseline without TrajMod places it at Rank 3. This demonstrates that TrajMod effectively leverages the trajectory-aware spatial context embedded in the text description.

F.3 Failure Case Analysis

We identify three dominant failure modes and present representative examples for each. Understanding these modes clarifies the practical boundary of TrajMod and motivates future work.

Fig. S2: Dataset sample visualization. Each row contains a satellite image with trajectory overlay, sequential street-view observations, and the corresponding route description.

Failure Mode 1: Video ambiguity. Figure S5 shows a case where the input video frames depict a straight road through dense vegetation with no distinctive landmarks. With TrajMod, the ground truth improves from Rank 1111 to Rank 338, but remains far from the top results. All top-ranked candidates exhibit similar road-and-tree patterns. This failure mode is common for trajectories through visually monotonous environments (rural roads, highways) where neither appearance nor trajectory geometry provides sufficient discriminative cues.

Failure Mode 2: Route description ambiguity. Figure S6 illustrates a case where the route description contains rich landmark mentions (e.g., Arlington Animal Hospital, Bank of Bennington, American flags), yet the described scene transitions through multiple visually similar residential and commercial zones. With TrajMod, the ground truth is ranked at position 18—a substantial improvement over the baseline (Rank 203), but still outside the top results. The top-ranked candidates share similar visual characteristics with the ground truth (two-lane roads, trees, small commercial buildings), suggesting that when named land-

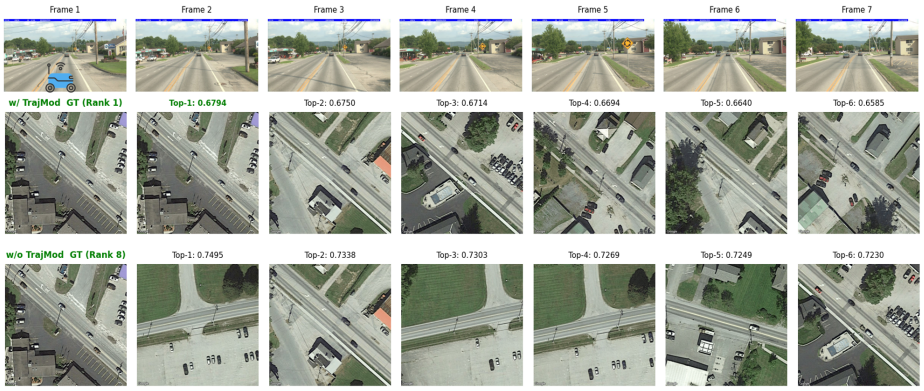

Fig. S3: Successful Cross-view video geo-localization. The top row shows the input video frames along the trajectory. The second and third rows show the top-6 matched satellite images with and without TrajMod, respectively. Green labels indicate the ground-truth tile and its similarity score. With TrajMod, the ground truth is correctly localized at Rank 1 (vs. Rank 8 without).

marks cannot be visually resolved from satellite imagery, text-to-satellite matching degrades to coarse scene-type matching.

Discussion: Failure modes and method boundary. The failure cases above reveal two complementary limitations: (i) when the visual scene is monotonous, TrajMod's geometric signal helps but cannot overcome the fundamental lack of appearance diversity (Failure Mode 1); (ii) when textual descriptions reference landmarks that are unresolvable from satellite imagery, the text encoder's semantic richness cannot compensate for the cross-view domain gap (Failure Mode 2). We additionally note that heading perturbation (Sec. D.5) constitutes a third failure mode where corrupted geometric metadata directly degrades TrajMod's effectiveness. These observations suggest that future improvements should focus on uncertainty modeling and on introducing additional modalities as priors to narrow the search space.

References

1. 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)
2. 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)
3. Wang, T., Zheng, Z., Sun, Y., Yan, C., Yang, Y., Chua, T.S.: Multiple-environment self-adaptive network for aerial-view geo-localization. Pattern Recognition **152**, 110363 (2024)

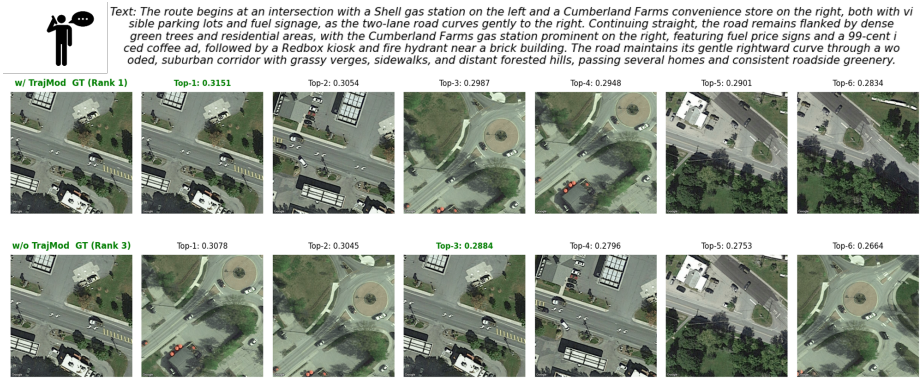

Fig. S4: Successful Cross-view text geo-localization. The input is a natural language route description (shown at the top). The top and bottom rows show the top-6 matched satellite tiles with and without TrajMod. Green labels indicate the ground truth. TrajMod localizes the correct tile at Rank 1, while the baseline ranks it at Rank 3.

4. 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)
5. 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)

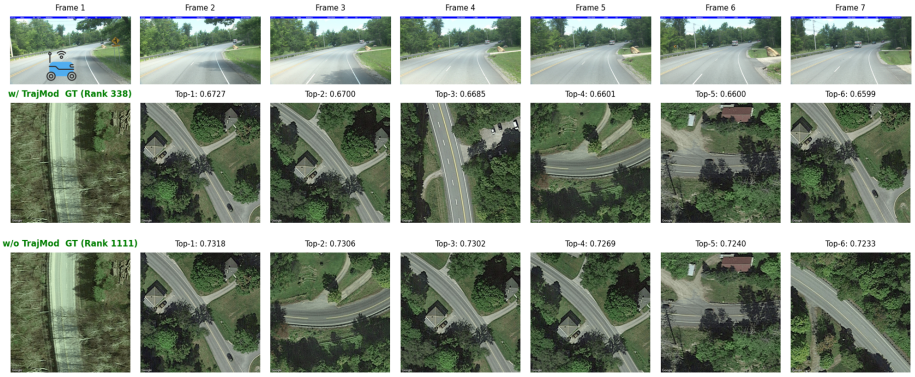

Fig. S5: Failure Mode 1: Video ambiguity. The trajectory passes through a visually monotonous rural road with dense vegetation, making the top-ranked satellite tiles nearly indistinguishable from the ground truth.

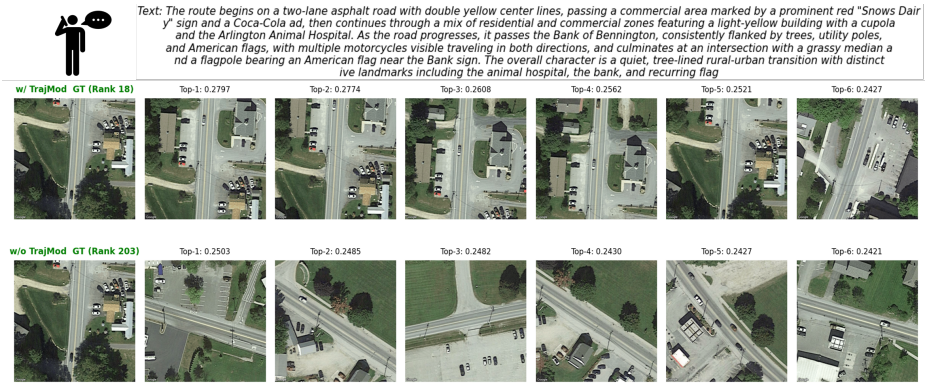

Fig. S6: Failure Mode 2: Route description ambiguity. The route passes through visually homogeneous residential areas, making it difficult to distinguish among satellite tiles that share similar road layouts, vegetation, and building patterns.