GEO-EMBED: TOWARDS UNIFIED MULTIMODAL EMBEDDINGS FOR URBAN UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Geospatial and urban applications increasingly require models to compare heterogeneous evidence across street-view imagery, remote-sensing observations, text descriptions, region proposals, and temporal change cues. However, existing multimodal embedding models and benchmarks are still largely designed and evaluated around general-purpose image-text matching, leaving unclear whether unified embedding space can support heterogeneous geospatial tasks involving spatial relationships, fine-grained semantics, and temporal changes. To address this gap, we make three key contributions. First, we introduce **GeoMEB**, a large-scale multimodal embedding benchmark that standardizes 45 urban evaluation tasks across retrieval, visual question answering, change detection, classification, and visual grounding, together with training collections comprising 1.32M examples and 286K evaluation queries. Second, we present **Geo-Embed**, a unified embedding model that adapts a shared vision-language backbone to encode heterogeneous geospatial inputs—including single or multiple images, text, regions, and masks—into a fixed-dimensional space for instruction-conditioned query-target matching. On GeoMEB, Geo-Embed achieves the strongest overall performance among representative multimodal embedders, with a 10.4% relative improvement over the strongest baseline. These results motivate future geospatial embedders that organize training and evaluation around explicit query-target relations, including semantic, cross-view, region-level, and temporal correspondence.

1 INTRODUCTION

Embedding models are becoming a core interface for geospatial and urban intelligence because many city-scale applications depend on comparing heterogeneous evidence about places (Brown et al., 2025; Liu et al., 2024; Zhang et al., 2024). Urban search, geo-localization (Li et al., 2026b), disaster monitoring, and socioeconomic sensing all require matching visual, textual, spatial, or temporal observations to task-relevant targets (Lu et al., 2018; Wang et al., 2024; Chen & Shi, 2020; Liu et al., 2025). This shared comparison problem makes embedding quality central to scalable urban analysis. Although these applications differ in form, they require urban queries and candidate targets to be represented in a shared space where task-relevant similarity can be computed efficiently.

Despite rapid progress in CLIP-style dual encoders and recent VLM-based embedders (Radford et al., 2021; Jiang et al., 2024; Zhang et al., 2025a; Li et al., 2026a), their ability to serve as a unified interface for urban intelligence remains unclear. First, existing embedding benchmarks are still dominated by generic image-text retrieval, general multimodal reasoning, or mixed-domain retrieval settings (Meng et al., 2025; Huang et al., 2026; Xiao et al., 2025), and they do not systematically organize urban problems under a unified, task-aware common embedding protocol. Second, urban similarity is relation-specific rather than purely semantic: cross-view geo-localization depends on viewpoint alignment (Zhu et al., 2021; Jia et al., 2024), region grounding depends on spatial granularity and region correspondence (Ertler et al., 2020; Li et al., 2024), and disaster monitoring depends on temporal change sensitivity (Alcantarilla et al., 2018; Chen & Shi, 2020; Huang et al., 2025). As a result, current models and evaluation resources provide limited evidence on whether a single representation space can support heterogeneous urban query-target relations across modalities, viewpoints, regions, and temporal states.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address these gaps, we introduce **GeoMEB**, a geospatial multimodal embedding benchmark that standardizes 45 tasks under a unified instruction-conditioned ranking protocol. GeoMEB covers retrieval, visual question answering, change detection, classification, and visual grounding, and casts each task as matching an instructed query to candidates from a task-specific pool. The benchmark includes 1.32M training examples and 286K evaluation queries across image, text, interleaved image-text, multi-image, region, and mask inputs (Figure 1). We further present **Geo-Embed**, a unified embedding model for geospatial multimodal data. Geo-Embed encodes queries and targets with side-specific instructions using a shared vision-language backbone, and trains the resulting representations with contrastive supervision over heterogeneous geospatial pairs.

Experiments on GeoMEB show that Geo-Embed achieves the strongest overall performance among representative CLIP-family models, instruction-tuned VLMs, and embedding-specialized VLMs. More importantly, the results suggest that future geospatial embedders should organize training and evaluation around explicit query-target relations, as semantic matching, cross-view alignment, region grounding, and temporal change matching respond differently to data composition and task mixing. Our main contributions are as follows:

- We introduce GeoMEB, a 45-task geospatial multimodal embedding benchmark with a unified instruction-conditioned ranking protocol.
- We develop Geo-Embed, an instruction-conditioned VLM embedding model for heterogeneous geospatial query-target matching.
- We provide a systematic study of multimodal embedders on GeoMEB, covering baseline comparisons, ablations, instruction sensitivity, and representation analyses.

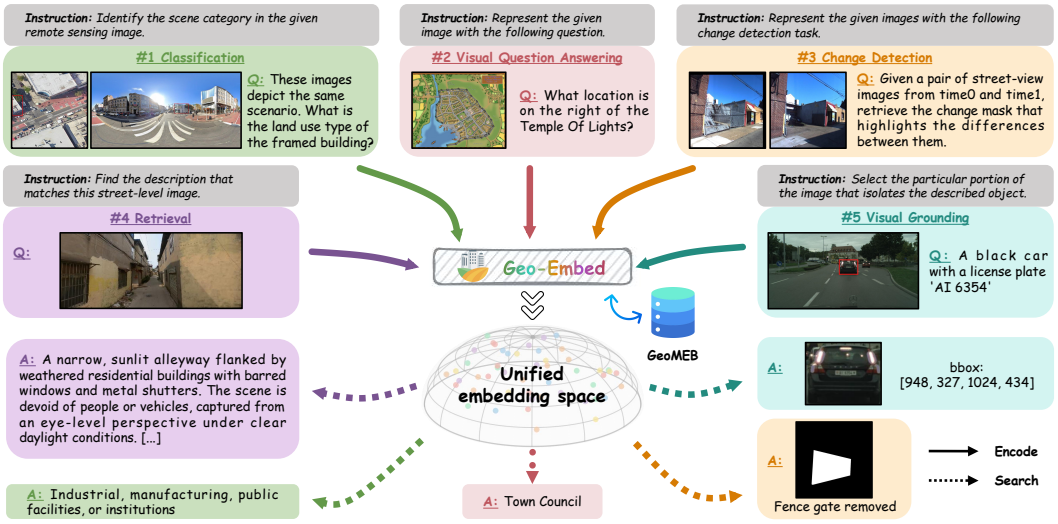


Figure 1: Overview of GeoMEB and Geo-Embed. GeoMEB unifies classification, visual question answering, change detection, retrieval, and visual grounding as instruction-conditioned embedding tasks; Geo-Embed encodes heterogeneous urban queries and target candidates into a shared representation space, enabling efficient similarity-based search across text, single- and multi-image inputs, regions, and masks under a common query-target matching framework.

2 Related Work

2.1 Multimodal Embedding Models

Multimodal embedding models map inputs from different modalities into a shared representation space for efficient retrieval. Existing approaches use either modality-specific dual encoders or VLM-based embedders that process interleaved inputs and task instructions. CLIP-style dual encoders learn image-text alignment with a contrastive objective and permit target candidate embeddings to be precomputed (Radford et al., 2021; Tschannen et al., 2025). However, fixed towers are less

suited to interleaved inputs or instructions that redefine the desired relation. VLM2Vec exemplifies the second paradigm by appending task instructions, using the final-token hidden state as the sequence representation, and optimizing it with InfoNCE (Jiang et al., 2024; van den Oord et al., 2019). Data-centric methods scale paired supervision for universal retrieval (Zhou et al., 2025c). Other systems broaden task and input coverage to videos, visual documents, audio, and universal multimodal retrieval (Zhang et al., 2025a; Meng et al., 2025; Xu et al., 2025). Recent native multimodal embedders further integrate retrieval and ranking within stronger VLM backbones (Li et al., 2026a; Shanbhogue et al., 2026). Reasoning-oriented approaches generate context or optimize retrieval behavior before producing embeddings (Cui et al., 2025; Lan et al., 2025; Jiang et al., 2026).

2.2 Geospatial and urban representation learning

Urban applications such as cross-view localization, temporal change analysis, and large-scale mapping require representations that remain informative across sensors, viewpoints, spatial scales, and time. Contrastive remote-sensing models align overhead imagery with language supervision at scale (Liu et al., 2024; Zhang et al., 2024; Wang et al., 2024). VLM-based approaches extend this alignment to grounded instructions and interleaved geospatial inputs (Kuckreja et al., 2024). AlphaEarth Foundations instead learns a global geospatial embedding field for large-scale mapping (Brown et al., 2025). These advances provide strong representations for remote-sensing retrieval, grounded understanding, and global mapping. Urban settings, however, require a common representation to relate overhead and street-level observations, geographic scales, temporal states, regions, and urban semantics. Existing geospatial methods typically cover only a subset of these relations within separate task formulations.

2.3 Multimodal embedding benchmarks

Multimodal embedding benchmarks evaluate whether shared representations generalize across tasks, input modalities, and reasoning requirements. MMEB provides a broad suite of multimodal retrieval tasks (Jiang et al., 2024). Its extensions add videos and visual documents (Meng et al., 2025), while MMEB-V3 expands evaluation to audio and agent-centric scenarios (Huang et al., 2026). MIEB broadens image embedding evaluation across task categories and languages (Xiao et al., 2025). MR^2-Bench instead emphasizes reasoning-intensive query-target relations (Zhou et al., 2025b). City-scale analysis depends on linking observations of the same place across platforms, spatial scales, and time, as well as connecting visual evidence with urban conditions. However, existing benchmarks emphasize general multimodal capabilities and do not jointly assess city-specific relations involving cross-view correspondence, geographic scale, temporal change, region grounding, and social economic indicators. GeoMEB fills this gap by standardizing these relations under a single instruction-conditioned ranking protocol.

3 GeoMEB: A multimodal embedding benchmark for massive urban tasks

3.1 Benchmark overview

We introduce the **Geospatial Multimodal Embedding Benchmark (GeoMEB)**, which standardizes 45 evaluation tasks across five meta-tasks: retrieval, visual question answering, change detection, classification, and visual grounding. These tasks span 11 fine-grained task types and cover cross-view, street-view, remote-sensing, and document inputs (Figure 2). GeoMEB contains 286K evaluation queries, while the training set provides 1.32 million examples. Table 1 lists dataset families, task settings, query-target modalities, and the training, query, and candidate pools sizes.

3.2 Benchmark dimensions

Task coverage. Tasks include matching street-view and satellite images, retrieving image captions, answering questions about scenes, grounding referred objects, detecting changes from image pairs, and selecting labels for urban perception and socioeconomic indicators. They are grouped by query-target relation, spanning image-text matching, cross-view localization, spatial reasoning, temporal comparison, region grounding, and score-label choosing. Definitions are provided in Appendix A.

Table 1: GeoMEB dataset and task statistics grouped by meta-task and dataset family. Arrows denote query-to-target modality direction, and ↔ denotes bidirectional cross-modal retrieval.

Meta-task	Dataset (#tasks)	Task type	Setting	Query → Target	#Train	#Queries	#Candidates
Retrieval (13 tasks)	RSICD Lu et al. (2018) (2)	ICM	Remote sensing (RS)	image ↔ text	43,670	1,093∼4,972	1,093∼4,972
	UAV-GeoLoc Wu et al. (2025) (2)	CG	Cross-view	image → image	440,277	33,346∼44,832	11,684∼24,922
	VIGOR Zhu et al. (2021) (1)	CG	Cross-view	image+text → image	3,045	2,693	2,568
	SV Localization Li et al. (2026c) (2)	ICM	Street view (SV)	image ↔ text	160,037	39,887	39,887
	Im2GPS3k Vo et al. (2017) (4)	CG; ICM	Street view	image ↔ text	–	1,141∼2,997	1,141∼2,997
	VRSBench Li et al. (2024) (2)	ICM	Remote sensing	image ↔ text	157,674	9,350	9,350
VQA (13 tasks)	VRSBench Li et al. (2024) (1)	MVU	Remote sensing	image+text → text	157,674	37,409	2,373
	MME-RealWorld Zhang et al. (2025b) (6)	MVU	Street view; Cross-view; Document; Remote sensing	image+text → text	23,620	668∼1,293	5
	UrBench Zhou et al. (2025a) (6)	USR; UER	Remote sensing; Street view; Cross-view	image+text → text	9,274	50∼225	2∼900
Grounding (3 tasks)	Cityscapes Cordts et al. (2016) (1)	SOG	Street view	image+text → image	14,868	2,500	2,500
	Mapillary Ertler et al. (2020) (1)	SOG	Street view	image+text → image	61,503	8,924	8,924
	VRSBench Li et al. (2024) (1)	ROG	Remote sensing	image+text → image	157,674	16,159	16,159
Change Detection (4 tasks)	LEVIR-CD Chen & Shi (2020) (1)	CMM	Remote sensing	image → image	509	128	128
	SYSU-CD Shi et al. (2022) (1)	CMM	Remote sensing	image+text → image	16,000	4,000	4,000
	VL-CMU-CD Alcantarilla et al. (2018) (1)	CMM	Street view	image+text → image	3,732	429	429
	OSF Recovery Huang et al. (2025) (1)	DDA	Street view	image → text	190	137	4
Classification (12 tasks)	AID Xia et al. (2017) (1)	LCC	Remote sensing	image → text	8,000	2,000	30
	PlacePulse Salesses (2012) (6)	SBC	Street view	image → text	40,640	654∼2,947	14∼18
	CityLens Liu et al. (2025) (4)	SBC	Cross-view	image → text	98,524	1,000	20
	UrBench Zhou et al. (2025a) (1)	LCC	Cross-view	image+text → text	9,274	120	12

Urban-view and temporal coverage. GeoMEB combines street-level imagery, unmanned aerial vehicle imagery, satellite imagery, and urban documents such as maps. Cross-view tasks connect ground-level scenes with imagery captured at different viewpoints and spatial scales, while change tasks compare observations of the same or related locations over time. GeoMEB therefore evaluates variation in viewpoint, geographic extent, sensor source, and observation time rather than treating urban imagery as a single visual domain.

Modality and target diversity. Queries may contain text, single or multiple images, or interleaved image-text content. Targets may consist of captions, answer choices, class labels, geolocations, images, cropped regions, or change masks (see Appendix C for task-level examples).

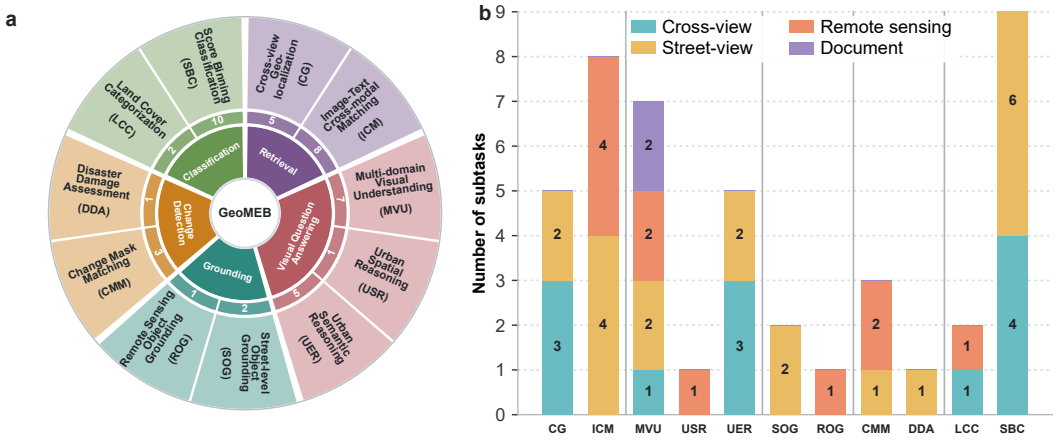


Figure 2: GeoMEB task taxonomy and input composition. (a) The 45 evaluation tasks grouped into five meta-tasks and 11 task types. (b) Their distribution across four modalities.

3.3 META-TASK AND EVALUATION DESIGN

GeoMEB reformulates every task as ranking over a task-specific candidate set. For task r, a model encodes an instruction-conditioned query q and each candidate $c \in \mathcal{C}_r$ into normalized embeddings $\mathbf{z}_q$ and $\mathbf{z}_c$, then ranks candidates by cosine similarity,

$$\mathrm{sim}(q, c) = \mathbf{z}_q^\top \mathbf{z}_c, \qquad c \in \mathcal{C}_r. \tag{1}$$

Task standardization. ***Retrieval*** tasks rank candidate images or texts for a query image, text, or image-text input. ***Visual Question Answering*** is converted into ranking over the answer vocabulary

or multiple-choice options. ***Classification*** uses textual class names or discretized score descriptions as candidates. ***Grounding*** ranks candidate regions against an image-text query, while ***Change Detection*** ranks masks or damage-status labels against multi-temporal observations.

Instruction and split standardization. Separate query and candidate instructions specify the relation being evaluated and the representation role of each input. We use the same task templates during training and evaluation. When official splits are available, we preserve them; otherwise, we construct random train/evaluation splits while ensuring that no target instances are shared across splits. Task definitions, dataset transformations, and instruction templates are detailed in Appendices A, B, and C, respectively.

4 GEO-EMBED: UNIFIED EMBEDDING MODEL FOR MASSIVE URBAN TASKS

Using a shared vision-language backbone and side-specific instructions, Geo-Embed maps heterogeneous geospatial queries and candidates into a shared embedding space for task-aware matching. Figure 3 illustrates its two-sided encoding architecture and contrastive training objective.

4.1 MODEL ARCHITECTURE

We first define how the query and candidate inputs are represented. Let $r \in \mathcal{T}$ denote an urban task with natural-language instructions p_r^q and p_r^c for the query and candidate sides, respectively. We decompose a query as $q_i = (q_i^{\text{vis}}, q_i^{\text{txt}})$ and a candidate as $c_j = (c_j^{\text{vis}}, c_j^{\text{txt}})$, where either modality component may also be empty. Images are converted into visual tokens by the visual tokenizer of the backbone, while textual descriptions and structured urban attributes are serialized as text tokens. We construct the two input sequences as

$$\widetilde{q}_i = [\text{Vis}(q_i^{\text{vis}}); \texttt{Instruct}: p_r^q; \texttt{Query}: q_i^{\text{txt}}],$$
$$\widetilde{c}_j = [\text{Vis}(c_j^{\text{vis}}); \texttt{Instruct}: p_r^c; \texttt{Candidate}: c_j^{\text{txt}}]. \tag{2}$$

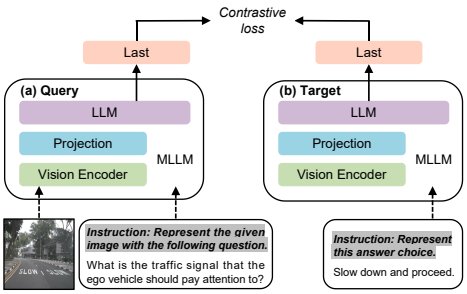


Figure 3: Architecture of Geo-Embed. The query and its target candidate are paired with side-specific instructions and independently encoded by the shared MLLM. Final-token hidden states from the last layer form the two representations, which are aligned through contrastive learning.

The same serialization covers text-only, image-only, image-text, and multi-image inputs by omitting empty components. Both sides are explicitly instruction-conditioned: p_r^q expresses the requested relation, whereas p_r^c anchors the representation role of an image, region, caption, answer, or label. This allows the same raw content to receive different embeddings across different semantic roles or tasks.

Geo-Embed uses a shared vision-language backbone F_θ to encode both sides. We take the last-layer hidden representation of the final token as the sequence representation:

$$\mathbf{h}_{q_i} = F_\theta(\widetilde{q}_i)^{(L)}_{\text{last}}, \qquad \mathbf{h}_{c_j} = F_\theta(\widetilde{c}_j)^{(L)}_{\text{last}}. \tag{3}$$

4.2 CONTRASTIVE TRAINING

For each training instance from task r, we pair an instruction-conditioned query (p_r^q, q_i) with one relevant instruction-conditioned candidate (p_r^c, c_i^+). The remaining candidates in the same batch provide negative supervision and are likewise encoded with their assigned candidate instructions. Let $\mathcal{N}_i$ denote the in-batch negative set for q_i, consisting of candidates paired with other queries in the batch. Thus, each query is optimized against one positive and multiple negatives. We use cosine similarity with temperature scaling,

$$s(q_i, c) = \frac{\text{sim}(q_i, c)}{\tau}, \tag{4}$$

where τ is a temperature hyperparameter. And then we train Geo-Embed with the InfoNCE objective,

$$\mathcal{L}_i = -\log \frac{\exp\left(s(q_i, c_i^+)\right)}{\exp\left(s(q_i, c_i^+)\right) + \sum_{c_j^- \in \mathcal{N}_i} \exp\left(s(q_i, c_j^-)\right)}. \tag{5}$$

For batch $\mathcal{B}$, the final objective averages the per-query losses:

$$\mathcal{L}_{\text{con}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathcal{L}_i. \tag{6}$$

This objective increases the similarity of each matched query-candidate pair relative to its in-batch negatives without encoding additional candidates.

4.3 Interleaved sub-batch sampling

To improve contrastive learning over heterogeneous geospatial data, we adopt an interleaved sub-batch sampling strategy. Let a training batch $\mathcal{B}$ of size B be partitioned into M sub-batches of size $b = B/M$. For each sub-batch, we first sample an input-modality group g_m, and then draw all samples in that sub-batch from the corresponding data pool $\mathcal{D}_{g_m}$:

$$\mathcal{B} = \bigcup_{m=1}^{M} \mathcal{B}_m, \qquad \mathcal{B}_m = \{(q_i, c_i)\}_{i=1}^{b} \sim \mathcal{D}_{g_m}^{b}. \tag{7}$$

Samples within each sub-batch share the same modality group, creating more similar and harder in-batch negatives, while different sub-batches preserve modality diversity within the full batch.

5 Experiments and Results

5.1 Experiment setting

Geo-Embed is initialized from Qwen3-VL-Embedding model and trained with the above instruction-conditioned contrastive objective using LoRA rank 16 and a batch size of 1024. All tasks are serialized into a unified query-target ranking format. Detailed hyperparameters, hardware, and implementation details are provided in Appendix E.

Baselines. We compare with three representative baseline model groups: CLIP-family models, instruction-tuned VLMs, and embedding-specialized VLMs.

Metrics. All tasks are evaluated as ranking over task-specific candidate pools. Image-text cross-modal retrieval tasks (ICM) use the mean of Recall@1, @5, and @10, since a query may have multiple semantically valid matches and retrieval quality is therefore better reflected across multiple ranking depths. Other tasks, such as classification, VQA, and change detection, typically have a single target answer and are evaluated using Recall@1. Candidate-pool size is treated as part of each task's native retrieval setting rather than normalized away. Table 2 reports the overall performance together with averages for fine-grained task types and meta-task groups, with detailed task-level results provided in Appendix D.

5.2 Main results

Geo-Embed achieves the strongest overall performance on GeoMEB. Geo-Embed achieves the highest overall average score of 25.48, outperforming Qwen3-VL-Embedding-8B (23.09), Qwen3-VL-Embedding-2B (21.60), and Seed1.6-Embedding (21.44). Geo-Embed ranks first on classification, retrieval, and grounding, and remains second-best on both VQA and change detection. These gains indicate that urban contrastive fine-tuning improves task-specific ranking performance across heterogeneous query-target relations. Figure 4 further illustrates this geometric shift: Qwen3-VL-Instruct-2B forms highly compact clusters because it is not designed for retrieval, making nearest-neighbor retrieval difficult; Qwen3-VL-Embedding-2B produces a more dispersed retrieval-oriented space; and fine-tuning brings most query-target pairs closer, although some tasks remain imperfectly aligned after training.

Table 2: Performance comparison on GeoMEB. LCC: land-cover categorization; SBC: score-binning classification; ICM: image-text cross-modal matching; CG: cross-view geo-localization; MVU: multi-domain visual understanding; USR: urban spatial reasoning; UER: urban semantic reasoning; SOG: street-level object grounding; ROG: remote-sensing object grounding; CMM: change mask matching; DDA: disaster damage assessment. The final "All" column averages over all meta-tasks, treating missing ones as 0. Best values are bolded and second-best values are underlined.

		Classification			Retrieval			VQA				Grounding			Change Detection			All
Model	**Size**	**LCC**	**SBC**	**All**	**ICM**	**CG**	**All**	**MVU**	**USR**	**UER**	**All**	**SOG**	**ROG**	**All**	**CMM**	**DDA**	**All**	
#Tasks		2	10	12	8	5	13	7	1	5	13	2	1	3	3	1	4	45
CLIP-family models																		
CLIP	0.43B	47.73	4.35	11.58	29.31	4.12	19.62	7.37	20.45	11.86	10.10	1.25	4.16	2.22	0.78	–	0.58	8.82
OpenCLIP	0.99B	41.91	4.88	11.06	40.06	7.99	27.73	11.69	27.27	11.20	12.70	1.66	5.41	2.91	0.26	–	0.20	10.92
RemoteCLIP	0.43B	57.54	1.72	11.02	26.00	0.24	16.10	9.39	7.95	10.62	9.75	0.74	0.95	0.81	0.52	–	0.39	7.61
GeoRSCLIP	0.99B	43.73	7.01	13.13	37.58	8.05	26.22	9.47	18.18	11.66	10.98	1.51	5.36	2.79	0.52	–	0.39	10.70
SkyScript-CLIP	0.43B	47.04	4.79	11.83	30.72	6.77	21.51	7.35	19.32	12.15	10.11	1.16	5.58	2.63	1.04	–	0.78	9.37
SigLIP2	0.88B	36.30	6.09	11.13	31.94	3.62	21.05	6.25	20.45	16.13	11.14	2.44	5.08	3.32	0.26	–	0.20	9.37
Instruction-tuned VLMs																		
GeoChat	7B	15.54	**8.91**	10.02	0.95	0.37	0.73	11.90	27.27	11.75	13.03	0.03	0.53	0.19	0.26	–	0.20	4.83
Qwen3-VL-Instruct	2B	10.96	3.74	4.94	0.30	0.03	0.20	20.78	1.14	12.96	16.26	0.04	0.01	0.03	0.08	41.61	10.46	6.38
Qwen3-VL-Instruct	8B	6.38	4.96	5.19	0.59	0.18	0.43	21.54	15.91	13.30	17.94	0.05	0.59	0.23	0.35	33.58	8.65	6.49
Embedding-specialized VLMs																		
UME-R1	7B	50.98	5.71	13.25	31.11	4.98	21.06	48.26	22.73	26.58	37.96	4.55	4.17	4.43	0.19	37.96	9.63	17.27
VLM2Vec-V2.0	3B	50.50	6.10	13.50	55.79	4.24	35.96	25.12	2.27	15.48	19.65	3.25	5.85	4.12	0.52	29.93	7.87	16.22
OmniEmbed-Nemotron	3B	33.98	6.43	11.02	36.29	2.87	23.44	14.15	11.36	11.75	13.01	1.96	2.81	2.24	0.65	34.31	9.07	11.76
Seed1.6-Embedding	N/A	56.48	7.25	15.46	56.70	6.82	37.52	42.27	43.18	24.37	35.45	6.38	7.68	6.81	1.81	**42.34**	**11.94**	21.44
Qwen3-VL-Embedding	2B	60.60	5.96	15.07	60.25	6.86	39.72	50.79	23.86	26.20	39.26	4.99	3.56	4.52	0.63	35.77	9.42	21.60
Qwen3-VL-Embedding	8B	54.12	3.99	12.35	63.31	8.76	42.33	**56.20**	**44.32**	30.07	**45.23**	7.02	6.27	6.77	1.11	31.68	8.75	23.09
Geo-Embed (Ours)	2B	**66.75**	7.25	**17.17**	**63.95**	**13.80**	**44.66**	49.96	39.77	**35.36**	43.56	**9.13**	**13.00**	**10.42**	**2.80**	37.96	11.59	**25.48**

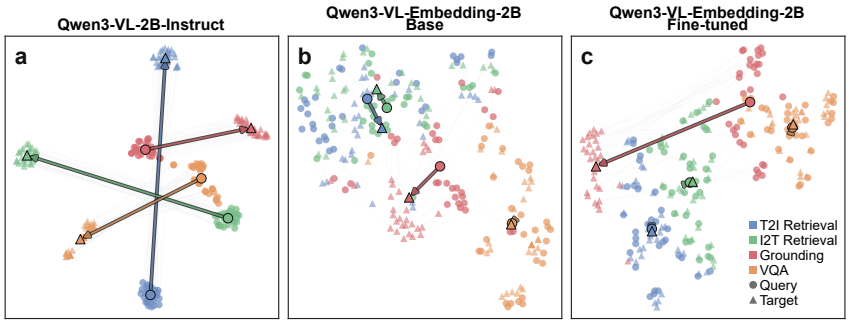


Figure 4: Feature space visualization: (a) Qwen3-VL-Instruct-2B, (b) Qwen3-VL-Embedding-2B, and (c) Qwen3-VL-Embedding-2B after VRSBench fine-tuning. Circles and triangles denote query and target embeddings, respectively.

Generic multimodal ability does not guarantee strong urban retrieval. CLIP-family models retain reasonable performance on some classic image-text and classification tasks, but their scores drop sharply on VQA, grounding, and change detection, where inputs are often interleaved, instruction-dependent, or region- and mask-oriented. Instruction-tuned VLMs show a different limitation: Qwen3-VL-Instruct and GeoChat can follow multimodal prompts, but their overall embedding scores remain low, indicating that instruction-following ability does not necessarily produce a suitable embedding space for retrieval. Even strong generic embedders are uneven across task types: Qwen3-VL-Embedding-8B achieves the best on VQA, and Seed1.6-Embedding is strongest on change detection. These results show that generic multimodal alignment alone is insufficient for GeoMEB, where models must support cross-view matching, urban semantics, region grounding, temporal comparison, and answer selection within one embedding interface.

Fine-grained results reveal urban tasks remain unsolved. Across models, ICM and LCC are the most mature tasks, corresponding to image-caption matching and land-cover or scene-label recognition. Several embedding-specialized models performed well on these tasks, and Geo-Embed further improves ICM to 63.95 and LCC to 66.75, suggesting that current embedders can already handle semantic correspondence when query and target share relatively direct visual or textual cues. VQA also benefits from embedding-specialized backbones, with Qwen3-VL-Embedding-8B and Geo-

Embed reaching 45.23 and 43.56 overall. In contrast, CG, SBC, SOG/ROG, and CMM remain much harder, covering cross-view geo-localization, urban perception indicator estimation, region grounding, and mask-level change matching. These gaps suggest that current multimodal embedders are useful for coarse semantic matching, but still struggle to align geospatial viewpoints, infer abstract city attributes, localize region-level targets, and compare fine-grained temporal changes.These task types therefore define concrete stress tests for future geospatial embedding models.

6 Ablation analysis

Effect of interleaved sub-batch sampling. We evaluate modality-aware interleaved sub-batching with different sub-batch sizes. Sub0 denotes fully random batching without grouping, while Sub32∼Sub256 denote modality-homogeneous sub-batches of the corresponding size.

Table 3 shows that larger modality-homogeneous sub-batches mainly benefit within-modality retrieval. Sub256 achieves the best scores on RS, SV, and Document inputs, improving over Sub0 by 0.85, 0.90, and 0.05 points, respectively. In contrast, Cross-view retrieval is best with Sub0 (29.78), while the grouped settings remain slightly lower (29.15∼29.67). This pattern suggests that homogeneous sub-batches strengthen discrimination within a modality, whereas cross-view retrieval benefits from the greater modality diversity preserved by random batching. Thus, the preferred sub-batch size depends on the target capability: Sub256 is most effective for within-modality retrieval, while Sub0 is better suited to cross-view alignment.

Table 3: Effect of modality-aware interleaved sub-batch sampling across input modalities.

Group	Sub0	Sub32	Sub64	Sub128	Sub256
RS	21.32	21.73	21.52	21.30	**22.17**
SV	27.59	27.33	28.18	27.84	**28.49**
Cross-view	**29.78**	29.67	29.15	29.26	29.43
Document	41.05	40.65	40.90	40.50	**41.10**

Task-modality interactions reveal compatibility and interference.

We next examine how training-data modality affects cross-task generalization. The training groups use comparable subsets from RS captioning (VRSBench, ∼40K), SV localization (CityLens, ∼40K), cross-view matching (VIGOR, ∼42K), and all-source combination (∼122K). Figure 5a shows clear modality-level interactions. RS-only training gives the best RS retrieval score of 35.0, and SV-only training gives the best SV retrieval score of 98.4, both exceeding the all-source setting on their closest evaluation tasks. In contrast, cross-view evaluation requires cross-view or all-source supervision: RS-only and SV-only training give very low cross-view scores, while the all-source setting reaches 21.4. This suggests that single-modality supervision sharpens within-modality retrieval, but cross-view matching requires cross-view alignment signals.

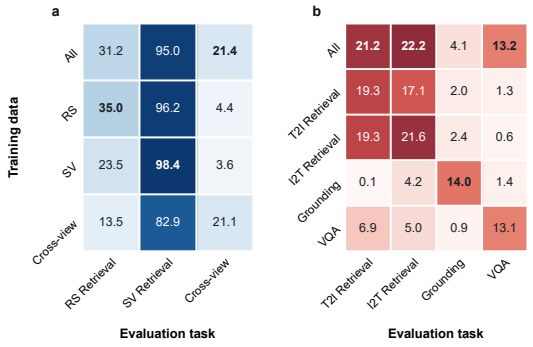


Figure 5: Performance comparison of models trained with different subsets. (a) compares remote-sensing, street-view, and cross-view subsets with the full mixture for retrieval; (b) compares retrieval, grounding, and VQA subsets with the full mixture across VRSBench task families.

On VRSBench, Figure 5b shows a similar interaction across task sources. All-source training gives the best T2I and I2T retrieval scores, reaching 21.2 and 22.2, and remains competitive on VQA with a score of 13.2. However, grounding behaves differently: grounding-only training reaches 14.0, whereas all-source training drops substantially. This indicates that multi-task supervision is not automatically additive: compatible tasks can reinforce shared semantic alignment, whereas structurally different tasks may interfere with specialized query-target relations. This is consistent with the feature space visualization in Figure 4: fine-tuning improves several query-target neighborhoods, but not all task relations become more compact, as shown by the enlarged VQA query-target separation.

Instruction-induced shifts are not uniformly target-oriented. We further test whether task instructions move query embeddings closer to their matched targets in the original embedding space. We annotate 50 aligned VIGOR triplets, each containing a street-view image, a remote-sensing image, and a neutral cross-view description. We report the direction-specific mean change in matched query-target cosine similarity (Figure 6); positive and negative values indicate movement toward and away from the matched target, respectively. The metric is formally defined in Appendix C.1.

Figure 6 shows that instruction-induced shifts do not translate into uniformly target-oriented movement. OmniEmbed-Nemotron-3B exhibits broadly positive shifts, with five of the six directed cross-modal settings moving closer to the matched target and the largest improvements appearing in text-to-visual directions, such as Text→RS (+0.0772) and Text→SV (+0.0719). However, this behavior is still asymmetric: SV→RS decreases alignment (-0.0390), even though the reverse direction RS→SV improves (+0.0393). Qwen3-VL-Embedding-2B shows the same direction-dependent pattern in a different form. Spatially grounded instructions improve several directions, especially SV→RS (+0.0643) and SV→Text (+0.0455), but still reduce alignment for Text→SV (-0.0448) and RS→SV (-0.0103). These results suggest that instruction-induced shifts interact unevenly with the pretrained embedding geometry rather than establishing a consistent bidirectional alignment between modality pairs. Future geospatial embedders should therefore not only respond to task instructions, but also align the induced movement with the intended spatial relation between query and target.

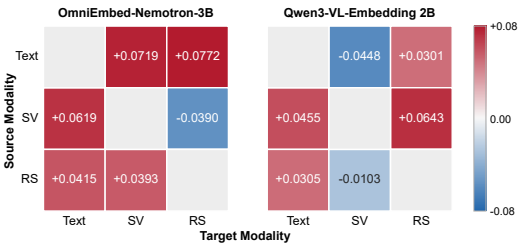


Figure 6: Instruction-induced alignment shifts, reporting the mean change in matched query-target cosine similarity when queries are encoded with or without task-specific instruction.

Hyperparameter sensitivity.

We identify two key parameters that affect Geo-Embed performance: LoRA rank and batch size. Experiments use a random 10% GeoMEB subset. Table 4 shows that LoRA rank has a non-monotonic effect: rank 16 gives the best overall average, rank 4 remains competitive and performs slightly better on classification and retrieval, while rank 32 degrades the average score. Increasing the batch size from 64 to 1024 further improves the overall score from 18.30 to 18.80, with the clearest gains on grounding and change detection. We therefore use LoRA rank 16 and batch size 1024 as the default setting.

Table 4: Hyperparameter sensitivity to LoRA rank and batch size.

Setting	Class.	Retr.	VQA	Ground.	ChgDet	Overall
LoRA rank comparison *(batch size = 64)*						
LoRA rank = 4	**15.28**	**39.58**	20.78	6.30	9.40	18.27
LoRA rank = 8	13.91	39.32	20.76	6.40	**9.59**	18.00
LoRA rank = 16	14.64	39.35	**20.99**	**7.17**	9.34	**18.30**
LoRA rank = 32	13.93	39.33	20.71	5.67	9.43	17.81
Batch size comparison *(LoRA rank = 16)*						
Batch size = 64	14.64	39.35	20.99	7.17	9.34	18.30
Batch size = 256	14.85	39.78	21.37	6.90	9.50	18.48
Batch size = 512	13.19	**40.07**	**21.77**	8.50	8.72	18.45
Batch size = 1024	14.92	39.94	20.26	**8.87**	**10.02**	**18.80**

7 Conclusion

This work reframes geospatial multimodal embedding as relation-centered query-target matching across heterogeneous urban evidence. We introduced GeoMEB, a 45-task benchmark that standardizes semantic matching, cross-view alignment, region-level grounding, and temporal change correspondence under a unified instruction-conditioned ranking protocol, and Geo-Embed, a unified embedding model trained with task-aware contrastive supervision. Experiments show that Geo-Embed achieves the strongest overall performance, while fine-grained analyses reveal that geospatial embedding quality depends strongly on the target relation. These results suggest that future geospatial embedders should make query-target relations explicit in training, sampling, instruction design, and evaluation. GeoMEB provides a foundation for building geospatial embedding models that are comparable across tasks and reusable across city-scale applications.

AI USE STATEMENT

In this work, we used generative AI tools to generate auxiliary textual annotations for VIGOR images used in the cross-view experiments described in the paper. Additionally, we used generative AI tools to draft sections of the manuscript and polish the writing for clarity and readability. We have not used generative AI tools for literature retrieval or discovery, research ideation or execution, or proving mathematical claims. No other uses beyond those disclosed in this statement apply to this work. We take responsibility for the final content of this work, including text, claims, annotations, and other research artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops GeoMEB and Geo-Embed using existing geospatial and urban datasets. These datasets may contain geographic, socioeconomic, and visual biases, and benchmark performance should not be interpreted as evidence of equal reliability across regions or communities. In particular, labels related to perceived safety, wealth, and other urban attributes reflect the definitions and annotation procedures of the source datasets rather than objective judgments about residents or communities. AI-generated textual annotations may also introduce inaccuracies or biases. Cross-view matching and geolocation capabilities carry potential privacy risks, including misuse for surveillance or identifying sensitive locations. We therefore emphasize responsible use and caution against relying on model outputs alone for consequential decisions affecting individuals or communities.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our work. Detailed descriptions of the benchmark construction, data processing procedures, model architecture, experimental settings, and evaluation protocols are provided in the main paper and appendices. Additional implementation details and ablation settings are also included to facilitate reproduction of our results. Upon completion of the review process, we will publicly release the code, associated data and model resources.

REFERENCES

Pablo F Alcantarilla, Simon Stent, German Ros, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42(7):1301–1322, 2018.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, 2025.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025.

ByteDance Seed Team. Built on seed1.6-flash, seed-1.6-embedding launched. https://seed.bytedance.com/en/blog/built-on-seed1-6-flash-seed-1-6-embedding-launched, June 2025. Accessed: 2026-08-28.

Hao Chen and Zhenwei Shi. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sensing*, 12(10):1662, 2020. ISSN 2072-4292.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016.

Xuanming Cui, Jianpeng Cheng, Hong-You Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Taipeng Tian, Qi Guo, Ser-Nam Lim, Aashu Singh, and Xiangjun Fan. Think Then Embed: Generative Context Improves Multimodal Embedding. In *The Fourteenth International Conference on Learning Representations*, 2025.

Christian Ertler, Jerneja Mislej, Tobias Ollmann, Lorenzo Porzi, Gerhard Neuhold, and Yubin Kuang. The Mapillary Traffic Sign Dataset for Detection and Classification on a Global Scale. In *ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, pp. 68–84, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58591-4.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks, 2023. URL https://arxiv.org/abs/2309.17425.

Haohang Huang, Xuan Lu, Mingyi Su, Xuan Zhang, Ziyan Jiang, Ping Nie, Kai Zou, Tomas Pfister, Wenhu Chen, Wei Zhang, Xiaoyu Shen, and Rui Meng. MMEB-V3: Measuring the Performance Gaps of Omni-Modality Embedding Models, 2026.

Tianyuan Huang, Chad Zanocco, Zhecheng Wang, Jackelyn Hwang, and Ram Rajagopal. Built environment disparities are amplified during extreme weather recovery. *Nature*, 648(8093):349–356, 2025. ISSN 1476-4687.

Pengyue Jia, Yiding Liu, Xiaopeng Li, Xiangyu Zhao, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, and Dawei Yin. G3: An Effective and Adaptive Framework for Worldwide Geolocalization Using Large Multi-Modality Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. Embed-RL: Reinforcement Learning for Reasoning-Driven Multimodal Embeddings, 2026.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. In *The Thirteenth International Conference on Learning Representations*, 2024.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. UME-R1: Exploring Reasoning-Driven Generative Multimodal Embeddings. In *The Fourteenth International Conference on Learning Representations*, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking, 2026a.

Xiang Li, Jian Ding, and Mohamed Elhoseiny. VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding. *Advances in Neural Information Processing Systems*, 37:3229–3242, 2024.

Yong Li, Furong Jia, Dacheng Yin, Kang Rong, Fengyun Rao, Jing Lyu, and Fan Zhang. Reverse: Reinforcing evidence verification and search for agentic image geo-localization. *arXiv preprint arXiv:2605.26861*, 2026b.

Yong Li, Weiyu Zhang, Ling Dai, Jian Yang, Dacheng Yin, Sirun Li, Jing Lyu, Fengyun Rao, and Fan Zhang. Unified urban tuning: Co-enhancing satellite and street view reasoning with a progressive tuning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6302–6311, 2026c.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. ISSN 1558-0644.

Tianhui Liu, Hetian Pang, Xin Zhang, Tianjian Ouyang, Zhiyuan Zhang, Jie Feng, Yong Li, and Pan Hui. CityLens: Evaluating Large Vision-Language Models for Urban Socioeconomic Sensing. In *The Fourteenth International Conference on Learning Representations*, 2025.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2018. ISSN 1558-0644.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Raghuveer Thirukovalluru, Xuan Zhang, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing Multimodal Embedding for Videos, Images, and Visual Documents. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Mark Philip Salesses. *Place Pulse : Measuring the Collaborative Image of the City*. PhD thesis, Massachusetts Institute of Technology, 2012.

Madhuri Shanbhogue, Zhe Li, Shanfeng Zhang, Gustavo Hernández Ábrego, Shih-Cheng Huang, Aashi Jain, Daniel Salz, Sonam Goenka, Chaitra Hegde, Ji Ma, Feiyang Chen, Jiaxing Wu, Tanmaya Dabral, Babak Samari, Kevin Poulet, Daniel Cer, Kaifeng Chen, Paul Suganathan, Hui Hui, Jovan Andonov, Philippe Schlattner, Jay Han, Iftekhar Naim, Wing Lowe, Vladimir Pchelin, Albert Yang, Yi-Ting Chen, Zhongli Ding, Grace Zhang, Georg Heigold, Yichang Chen, Antoine Reveillon, Brendan Mccloskey, Wenlei Zhou, Dahun Kim, Rui Meng, Emma Wang, Jack Zheng, Halley Fede, Zhen Yang, Keegan Mosley, Brian Potetz, Sahil Dua, Henrique Schechter Vera, Shen Gao, Hesen Zhang, Andreas Hess, Hengxuan Ying, Alberto Montes, Karan Gill, Min Choi, Sebastian Russo, Anja Hauth, Jinhyuk Lee, Michael Boratko, Megan Barnes, Vikram Rao, Claudiu Musat, Cyril Allauzen, Ehsan Variani, Shankar Kumar, Tom Bagby, Junyi Jiao, Yang Gu, Tengxin Li, Ayush Agrawal, Roberto Santana, Dev Nath, Stephen Karukas, Shuoxuan Han, Lucia Loher, Alice Twu, Nidhi Vyas, Siddharth Bhai, Frank Palma Gomez, Wangyuan Zhang, Chaoren Liu, Jizheng Yang, Steve Qiu, Shijie Zhang, Sujay Kulkarni, Sascha Rothe, Sean Nakamoto, Raphael Hoffmann, Zach Gleicher, Yunhsuan Sung, Qin Yin, Tom Duerig, and Mojtaba Seyedhosseini. Gemini Embedding 2: A Native Multimodal Embedding Model from Gemini, 2026.

Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. A Deeply Supervised Attention Metric-Based Network and an Open Aerial Image Dataset for Remote Sensing Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. ISSN 1558-0644.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, 2019.

Nam Vo, Nathan Jacobs, and James Hays. Revisiting IM2GPS in the Deep Learning Era. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2621–2630, 2017.

Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. SkyScript: A Large and Semantically Diverse Vision-Language Dataset for Remote Sensing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5805–5813, 2024. ISSN 2374-3468.

Rouwan Wu, Jiacheng Deng, Mingyu Mou, Xingyi He, Maojun Zhang, Yu Liu, and Shen Yan. UAV-GeoLoc: A Large-Vocabulary Dataset and Geometry-Transformed Method for UAV Geo-Localization. *IEEE Robotics and Automation Letters*, 10(9):8978–8985, 2025. ISSN 2377-3766.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. ISSN 1558-0644.

Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. MIEB: Massive Image Embedding Benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22187–22198, 2025.

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. Omni-Embed-Nemotron: A Unified Multimodal Retrieval Model for Text, Image, Audio, and Video, 2025.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging Modalities: Improving Universal Multimodal Retrieval by Multimodal Large Language Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9274–9285, 2025a.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? *International Conference on Learning Representations*, 2025:89655–89701, 2025b.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. RS5M and GeoRSCLIP: A Large-Scale Vision- Language Dataset and a Large Vision-Language Model for Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024. ISSN 1558-0644.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(10):10707–10715, 2025a. ISSN 2374-3468.

Junjie Zhou, Ze Liu, Lei Xiong, Jin-Ge Yao, Yueze Wang, Shitao Xiao, Fenfen Lin, Miguel Hu Chen, Zhicheng Dou, Siqi Bao, Defu Lian, Yongping Xiong, and Zheng Liu. MR^2-Bench: Going Beyond Matching to Reasoning in Multimodal Retrieval, 2025b.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. MegaPairs: Massive Data Synthesis for Universal Multimodal Retrieval. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095, Vienna, Austria, 2025c. Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5316–5325, 2021.

APPENDIX

A TASK DEFINITIONS

GeoMEB contains 45 evaluation subtasks spanning retrieval, visual question answering, change detection, classification, and visual grounding. The 11 task types are cross-view geo-localization (CG),

image-text cross-modal matching (ICM), score binning classification (SBC), multi-domain visual understanding (MVU), urban spatial reasoning (USR), urban semantic reasoning (UER), street-level object grounding (SOG), remote-sensing object grounding (ROG), change mask matching (CMM), disaster damage assessment (DDA), and land-cover categorization (LCC). Complete GeoMEB inventory is summarized in Appendix Table 5.

GeoMEB converts each task into instruction-conditioned ranking over a task-specific target pool. The taxonomy contains five meta-tasks and eleven fine-grained task types:

Retrieval. Retrieval tasks ask the model to identify the matching item for a query from a large candidate pool. The query and target may use the same modality or different modalities, such as image-to-text (i2t) retrieval, text-to-image (t2i) retrieval, UAV-to-satellite matching, or street-view-to-satellite geo-localization. These tasks evaluate whether the embedding space preserves semantic and geospatial correspondence across viewpoint, scale, and representation changes.

- **Cross-view geo-localization (CG):** matches observations from different geographic views or representations, such as street-level images, satellite images, and location descriptions. It emphasizes viewpoint invariance and spatial correspondence.
- **Image-text cross-modal matching (ICM):** aligns images with textual descriptions in either direction. It emphasizes semantic compatibility between visual content and language.

Visual question answering. VQA tasks start from an image-conditioned question and are reformulated as answer retrieval. Each query contains visual evidence and a natural-language question, while the target pool contains candidate answers from the original answer vocabulary or multiple-choice set. This formulation tests whether an embedding model can represent not only visual content, but also the spatial, semantic, and contextual relation requested by the question.

- **Multi-domain visual understanding (MVU):** covers general visual understanding across remote sensing, street view, documents, and mixed urban scenes.
- **Urban spatial reasoning (USR):** focuses on spatial quantities and relations, such as counting, localization, and spatial comparison.
- **Urban semantic reasoning (UER):** covers scene-level semantic judgments, including object attributes, road understanding, role-based reasoning, traffic-sign interpretation, and visual-prompt reasoning.

Grounding. Grounding tasks require localizing the visual region described by an image-text query. Instead of generating bounding boxes or masks, we cast grounding as ranking candidate crops or regions and selecting the one that matches the referring expression. This setting evaluates fine-grained region-level alignment, where the correct target may depend on object attributes, spatial relations, and scene context.

- **Street-level object grounding (SOG):** grounds textual references in street-view imagery, where objects appear at human scale and under perspective distortion.
- **Remote-sensing object grounding (ROG):** grounds textual references in overhead imagery, where targets are often smaller, denser, and more scale-dependent.

Change detection. Change detection tasks compare observations of the same or related place across time. The query contain bi-temporal or multi-temporal imagery, sometimes with task text, and the candidate can be a change mask or a textual damage or recovery label. These tasks test whether embeddings capture temporal evidence and physical change rather than only static scene similarity.

- **Change mask matching (CMM):** matches image pairs to change masks or visual change targets, emphasizing differences rather than static scene content.
- **Disaster damage assessment (DDA):** maps multi-temporal observations to damage or recovery-status labels, testing semantic interpretation of physical change. DDA is unavailable for CLIP-family models and GeoChat, which do not support multi-image inputs.

Classification. Classification tasks convert recognition or scoring problems into ranking over class candidates. For scene recognition, candidates are class names; for urban indicators or

perception scores, candidates are discretized score labels. This design tests whether the embedding space can associate visual urban evidence with semantic categories and ordinal descriptions.

- **Land-cover categorization (LCC):** maps remote-sensing or cross-view visual evidence to land-cover or scene categories.
- **Score binning classification (SBC):** converts continuous or ordinal urban indicators into discrete textual targets, including perceived urban qualities and socioeconomic or built-environment indicators.

B DATASET DESCRIPTIONS

This section provides dataset descriptions for the GeoMEB inventory. Below we describe the original dataset sources and how each is used for specfic tasks.

AID. AID (Xia et al., 2017) is a dataset for aerial scene classification, designed to evaluate recognition over high-resolution remote-sensing scenes with diverse land-use and land-cover categories. GeoMEB uses AID for scene classification by converting the original 30-way recognition problem into image-to-text retrieval, where each image is matched against a candidate pool of scene-category labels.

RSICD. RSICD (Lu et al., 2018) is a remote-sensing image captioning dataset containing aerial images with sentence-level descriptions. GeoMEB uses these image-caption pairs for bidirectional image-text cross-modal matching: image-to-text retrieval treats the remote-sensing image as the query and captions as candidates, while text-to-image retrieval uses the caption as the query and retrieves the matching image.

UAV-GeoLoc. UAV-GeoLoc (Wu et al., 2025) targets UAV geo-localization, where aerial views captured from UAV platforms are matched to geo-tagged satellite references under viewpoint, scale, and geometric discrepancies. GeoMEB uses its country-level and terrain-level variants as two cross-view retrieval subtasks; in both cases, the query side contains the UAV view and the target side contains satellite candidates, testing whether an embedding model can bridge aerial-to-satellite appearance differences.

VIGOR. VIGOR (Zhu et al., 2021) is a cross-view geo-localization dataset designed to move beyond the idealized one-to-one matching assumption between street-view and satellite images. GeoMEB evaluates street-view-to-satellite retrieval on VIGOR, where each query combines street-level visual evidence with text context and the model retrieves the matching satellite image from a candidate pool, making the task a direct test of cross-view urban alignment.

CityLens. CityLens (Liu et al., 2025) benchmarks vision-language models for urban socioeconomic sensing by linking satellite and street-view observations with region-level urban indicators. GeoMEB uses four indicator-prediction subtasks—GDP, population density, average building height, and healthcare accessibility—where a query combines satellite imagery and multiple street-view images, and the target is a discretized normalized score label.

SV Localization. SV Localization is derived from the street-view data used in Unified Urban Tuning (Li et al., 2026c). It contains street-level imagery from both Google Street View (GSV) and Baidu Street View (BSV), covering 241 cities while each city provides roughly 850 images.

Im2GPS3k. Im2GPS3k (Vo et al., 2017) is a classic image geolocalization dataset with geographic metadata and brief latitude-longitude descriptions. GeoMEB uses it for zero-shot street-view geolocation via bidirectional GPS-region retrieval. We further annotate each image with a detailed scene description to construct bidirectional caption retrieval.

VRSBench. VRSBench (Li et al., 2024) is a versatile remote-sensing vision-language dataset that combines detailed captions, object references, and question-answer pairs for image understanding.

Table 5: Detailed GeoMEB task-level statistics.

Dataset	Subtask	Task type	Setting	Query → Target	#Train	#Queries	#Candidates
Retrieval (13 tasks)							
RSICD	i2t	ICM	RS	image → text	43,670	1,093	4,972
RSICD	t2i	ICM	RS	text → image	43,670	4,972	1,093
UAV-GeoLoc	country	CG	Cross-view	image → image	172,237	44,832	24,922
UAV-GeoLoc	terrain	CG	Cross-view	image → image	268,040	33,346	11,684
VIGOR	–	CG	Cross-view	image+text → image	3,045	2,693	2,568
SV Localization	i2t	ICM	SV	image → text	160,037	39,887	39,887
SV Localization	t2i	ICM	SV	text → image	160,037	39,887	39,887
Im2GPS3k	GPS i2t	CG	SV	image → text	–	2,997	1,141
Im2GPS3k	GPS t2i	CG	SV	text → image	–	1,141	2,997
Im2GPS3k	Caption i2t	ICM	SV	image → text	–	1,805	1,805
Im2GPS3k	Caption t2i	ICM	SV	text → image	–	1,805	1,805
VRSBench	caption i2t	ICM	RS	image → text	157,674	9,350	9,350
VRSBench	caption t2i	ICM	RS	text → image	157,674	9,350	9,350
Visual Question Answering (13 tasks)							
VRSBench	vqa	MVU	RS	image+text → text	157,674	37,409	2,373
MME-RealWorld	AutonomousDriving	MVU	SV	image+text → text	5,211	1,293	5
MME-RealWorld	MME-HD-CN	MVU	Cross-view	image+text → text	2,635	684	5
MME-RealWorld	diagram_and_table	MVU	Document	image+text → text	4,766	1,167	5
MME-RealWorld	monitoring_images	MVU	SV	image+text → text	2,824	668	5
MME-RealWorld	ocr_cc	MVU	Document	image+text → text	5,004	1,236	5
MME-RealWorld	remote_sensing	MVU	RS	image+text → text	3,180	858	5
UrBench	counting	USR	RS	image+text → text	351	88	7
UrBench	object-attribute-recognition oe	UER	Cross-view	image+text → text	340	85	15
UrBench	road-understanding oe	UER	SV	image+text → text	201	51	2
UrBench	role-based-reasoning	UER	Cross-view	image+text → text	898	225	900
UrBench	traffic-sign-reasoning	UER	SV	image+text → text	400	100	400
UrBench	visual-prompt-reasoning	UER	Cross-view	image+text → text	199	50	200
Grounding (3 tasks)							
Cityscapes	–	SOG	SV	image+text → image	14,868	2,500	2,500
Mapillary	–	SOG	SV	image+text → image	61,503	8,924	8,924
VRSBench	referring	ROG	RS	image+text → image	157,674	16,159	16,159
Change Detection (4 tasks)							
LEVIR-CD	–	CMM	RS	image → image	509	128	128
SYSU-CD	–	CMM	RS	image+text → image	16,000	4,000	4,000
VL-CMU-CD	–	CMM	SV	image+text → image	3,732	429	429
OSF Recovery	–	DDA	SV	image → text	190	137	4
Classification (12 tasks)							
AID	scene classification	LCC	RS	image → text	8,000	2,000	30
PlacePulse	safety	SBC	SV	image → text	10,995	2,754	14
PlacePulse	lively	SBC	SV	image → text	11,765	2,947	16
PlacePulse	beautiful	SBC	SV	image → text	6,052	1,520	16
PlacePulse	wealthy	SBC	SV	image → text	5,462	1,372	18
PlacePulse	depressing	SBC	SV	image → text	3,780	953	18
PlacePulse	boring	SBC	SV	image → text	2,586	654	18
CityLens	gdp	SBC	Cross-view	image → text	6,556	1,000	20
CityLens	pop	SBC	Cross-view	image → text	56,560	1,000	20
CityLens	height	SBC	Cross-view	image → text	7,056	1,000	20
CityLens	health	SBC	Cross-view	image → text	28,352	1,000	20
UrBench	scene-recognition	LCC	Cross-view	image+text → text	480	120	12
Training-only Collection (1 item)							
MP16Pro	–	–	Training-only	–	200,000	–	–

GeoMEB uses its caption retrieval, VQA, and referring-expression grounding components, covering image-text retrieval, image-question-to-answer matching, and image-expression-to-region retrieval within the same remote-sensing source.

UrBench. UrBench (Zhou et al., 2025a) is a dataset for evaluating large multimodal models in multi-view urban scenarios, with tasks designed around urban perception, spatial reasoning, and semantic understanding. GeoMEB uses its counting, object-attribute recognition, road understanding, role-based reasoning, traffic-sign reasoning, visual-prompt reasoning, and scene-recognition subtasks, covering VQA-style reasoning and classification under standardized ranking.

MME-RealWorld. MME-RealWorld (Zhang et al., 2025b) evaluates multimodal models on high-resolution real-world scenarios that are difficult even for humans, including visually dense and domain-diverse image-question settings. GeoMEB uses its autonomous driving, high-resolution Chinese imagery, diagram-and-table, monitoring-image, OCR-centered, and remote-sensing subsets as VQA tasks. Each task is converted into image-plus-question to answer-choice retrieval, with a small candidate pool of answer options.

Cityscapes. Cityscapes (Cordts et al., 2016) is a large-scale urban street-scene dataset for semantic and instance-level understanding, collected from stereo video sequences across diverse city environments. GeoMEB uses Cityscapes for street-level object grounding: the query consists of a street-view image and a referring expression, and the candidate pool contains cropped regions, requiring the model to retrieve the region that matches the described object.

Mapillary. The Mapillary Traffic Sign Dataset (Ertler et al., 2020) is a large-scale street-level traffic-sign dataset for detection and classification across diverse geographic, weather, and lighting conditions. GeoMEB uses Mapillary as a street-level grounding source focused on traffic-sign and street-scene regions. As with Cityscapes, GeoMEB evaluates image-plus-text queries against cropped image-region candidates.

LEVIR-CD. LEVIR-CD (Chen & Shi, 2020) is a remote-sensing change-detection dataset built from bi-temporal image pairs and designed to evaluate whether models can identify meaningful land-surface changes despite illumination and registration variation. GeoMEB reformulates it as change-mask matching: the query contains two images of the same location at different times, and the target is the corresponding change mask.

SYSU-CD. SYSU-CD (Shi et al., 2022) is another aerial-image change-detection dataset, introduced to study robust change representations under noise, scale variation, and pseudo-change interference.

VL-CMU-CD. VL-CMU-CD (Alcantarilla et al., 2018) extends change detection to street-view imagery captured over time, focusing on structural changes observed from vehicle-mounted camera sequences. GeoMEB uses VL-CMU-CD as a street-level change-detection task: the query contains a pair of street-view images from different times, and the candidate is a change mask that localizes visual changes.

OSF Recovery. OSF Recovery (Huang et al., 2025) studies post-disaster recovery by using street-level imagery to assess how built environments recover after extreme weather events. GeoMEB converts this task into retrieval over recovery-status labels, so the model must match multi-temporal visual evidence to the correct disaster-recovery category.

PlacePulse. PlacePulse (Salesses, 2012) captures collective human perception of urban street-view environments through crowd judgments of visual qualities such as safety and liveliness. GeoMEB uses six perception dimensions—safety, liveliness, beauty, wealth, depression, and boringness—and converts each continuous perception score into textual score labels at 0.5-point intervals for image-to-text ranking.

MP16Pro. MP16Pro (Jia et al., 2024) is released with G3 for worldwide geolocalization by augmenting MP16 with textual geographical descriptions associated with image locations. We uses MP16Pro only as auxiliary training set. It provides street-view geolocation-style supervision, helping the model learn location-aware urban representations.

C TASK FORMULATION AND EXAMPLES

GeoMEB uses an instruction-conditioned query-target framework for all task examples. Query-side instructions define the evaluated relation, while candidate-side instructions define the representation role of each target candidate. Appendix Table 6∼10 provide standardized formulations grouped by meta-task. Figure 7 shows a VIGOR triplet used for instruction-induced embedding shift analysis.

C.1 INSTRUCTION-SHIFT METRIC

We quantify instruction-induced movement for a directed source-target modality pair $m_s \rightarrow m_t$. For example i, let $\mathbf{z}_{q_i}^{\text{ins}}$ and $\mathbf{z}_{q_i}^{\text{noins}}$ denote the normalized query embeddings obtained with the task-specific instruction and without an instruction, respectively. Let $\mathbf{z}_{t_i}$ denote the matched target embedding obtained with the modality-default target instruction. We define the direction-level alignment shift as

$$\Delta_{\cos}(m_s \rightarrow m_t) = \frac{1}{N} \sum_{i=1}^{N} \left[(\mathbf{z}_{q_i}^{\text{ins}})^\top \mathbf{z}_{t_i} - (\mathbf{z}_{q_i}^{\text{noins}})^\top \mathbf{z}_{t_i} \right]. \tag{8}$$

Positive values indicate that the task instruction moves the query closer to its matched target, whereas negative values indicate movement away from it.

Table 6: Retrieval subtask examples. Each row shows one standardized retrieval formulation, with target-side fields shaded. In Query-text and Target-text columns, *italic text* denotes the encoding instruction and **bold text** denotes the query or target content; the same convention applies to the following tables.

Dataset	Subtask	Query text	Query image	Target text	Target image
RSICD	i2t	*Represent this remote sensing image for retrieval.*		*Find a remote sensing image matching this description.* **There are two big baseball fields on this earth.**	-
RSICD	t2i	*Find a remote sensing image matching this description.* **Several stones are on the beach.**	-	*Represent this remote sensing image.*	
UAV-GeoLoc	country	*Find the satellite image matching this UAV aerial view.*		*Represent the given satellite image.*	
UAV-GeoLoc	terrain	*Find the satellite image matching this UAV aerial view.*		*Represent the given satellite image.*	
VIGOR	cross-view retrieval	*Given the street view image and its description, identify the matching satellite image.*		*Represent the given satellite image for cross-view retrieval.*	
SV Localization	i2t	*Find the description that matches this street-level image.*	(See full caption below due to length limit.)	*Represent the given street-level scene description.* **(See full caption below due to length limit.)**	-
SV Localization	t2i	*Find the street-level image that matches this description.* **(See full caption below due to length limit.)**	-	*Represent the given street-level image.*	(See full caption below due to length limit.)
Im2GPS3k	GPS i2t	*Find the geographic region that best matches the location shown in this street-level photo.*		*Represent the given geographic region.* **A geographic region between latitude 40.7°N to 40.8°N and longitude 74.1°W to 74.0°W.**	-
Im2GPS3k	GPS t2i	*Find the street-level images taken within this geographic region.* **A geographic region between latitude 19.6°N to 19.7°N and longitude 156.0°W to 155.9°W.**	-	*Represent the given street-level image.*	
Im2GPS3k	Caption i2t	*Find the description that matches this street-level image.*	(See full caption below due to length limit.)	*Represent the given street-level scene description.* **(See full caption below due to length limit.)**	-
Im2GPS3k	Caption t2i	*Find the street-level image that matches this description.* **(See full caption below due to length limit.)**	-	*Represent the given street-level image.*	(See full caption below due to length limit.)

Table 6 (continued)

Dataset	Subtask	Query text	Query image	Target text	Target image
VRSBench	caption i2t	*Find a caption matching this remote sensing image.*		*Represent this remote sensing image caption.* **The high-resolution image displays a body of water with jetties extending into it. There is one small ship located in the top-middle area of the image, aligned with one of the jetties.**	-
VRSBench	caption t2i	*Find a remote sensing image matching this caption.* **The high-resolution image displays a pair of tennis courts surrounded by trees. The courts are adjacent to each other with their longer sides in parallel. The surrounding area includes a well-maintained grassy field to the south of the courts, and to the east, there is a small building.**	-	*Represent this remote sensing image.*	

Table 7: Visual question answering subtask examples.

Dataset	Subtask	Query text	Query image	Target text	Target image
VRSBench	vqa	*Represent the given remote sensing image with the following question.* **What color are the planes predominantly?**		*Represent this answer.* **White.**	-
MME-RealWorld	Autonomous Driving	*Represent the given image with the following question.* **This image shows the front view of the ego car. What is the future state of the brown suv in the middle?**		*Represent this answer choice.* **Brake gently to a stop.**	-
MME-RealWorld	MME-HD-CN	*Represent the given image with the following question.* **What is the larger font content on the external wall above the entrance of the building in the picture? (in Chinese)**		*Represent this answer choice.* **Library (in Chinese).**	-
MME-RealWorld	diagram-and-table	*Represent the given image with the following question.* **What is the revenue (USD) of Multi-Passenger Taxis in 2033 in the Revenue Breakdown table?**		*Represent this answer choice.* **1,354,686**	-
MME-RealWorld	monitoring-images	*Represent the given image with the following question.* **What material is the roof of the small white house on the roof of the building in the middle of the picture?**		*Represent this answer choice.* **iron**	-
MME-RealWorld	ocr-cc	*Represent the given image with the following question.* **What is the amusement project above the TarzaN SwiNGS on the right side of the cable car in the picture?**		*Represent this answer choice.* **CANNON BALL FALLS**	-
MME-RealWorld	remote-sensing	*Represent the given image with the following question.* **Where are the football fields located in this picture?**		*Represent this answer choice.* **Beside the water in the upper right area of this picture**	-
UrBench	counting	*Represent the given image with the following question.* **How many printed white or yellow crosswalks are shown in this image?**		*Represent the user's input.* **4**	-
UrBench	object-attribute-recognition oe	*Represent the given image with the following question.* **The satellite image depicting a specific area of a city. In this image, I have highlighted the rooftop of a building with a red box. The panoramic street view image is taken from the center of the area shown in the satellite image. This panoramic image covers a full 360-degree view. Your task is to identify the highlighted building in the panoramic street view image and determine the number of floors of this building based on its front or side view. To do this, follow these steps: Locate the building highlighted with the red box in the satellite image within the panoramic street view image. Analyze the building's front or side view in the street view image to determine its number of floors. Provide the number of floors of the building.**		*Represent the user's input.* **12**	-

Table 7 (continued)

Dataset	Subtask	Query text	Query image	Target text	Target image
UrBench	road-understanding oe	*Represent the given image with the following question.* **Are the following four images captured successively on the same road? If they are captured successively on the same road, then all four images will contain many of the same objects, such as buildings, but from slightly different viewpoints. Otherwise, no identical buildings would appear in any of the four images. Answer yes or no based on the hint above.**		*Represent the user's input.* **Yes**	-
UrBench	role-based-reasoning	*Represent the given image with the following question.* **If I am a pedestrian walking on this street, are there any stores that I can stop in and grab a quick bite to eat?**		*Represent the user's input.* **No, there are no restaurants or cafes on this street.**	-
UrBench	traffic-sign-reasoning	*Represent the given image with the following question.* **According to the traffic sign in box 0, what is the maximum speed limit on this road?**		*Represent the user's input.* **120 km/h**	-
UrBench	visual-prompt-reasoning	*Represent the given image with the following question.* **What is the person in the bounding box likely doing?**		*Represent the user's input.* **Crossing the street**	-

Table 8: Grounding subtask examples.

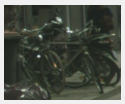
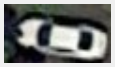
Dataset	Subtask	Query text	Query image	Target text	Target image
Cityscapes	street-level object grounding	*Represent the given street scene image with the described object.* **Where exactly is a bicycle located? Return the bounding box.**		*Represent the given cropped region of the street scene.*	
Mapillary	street-level object grounding	*Represent the given street scene image with the described object.* **Identify the position of the warning winding road first right g1 sign in this image.**		*Represent the given cropped region of the street scene.*	
VRSBench	referring	*Represent the given remote sensing image with the described object.* **The small white vehicle is parked on the right edge of the paved area.**		*Represent the given cropped region of the remote sensing image.*	

Table 9: Change-detection subtask examples.

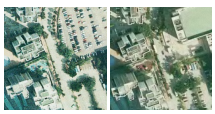
Dataset	Subtask	Query text	Query image	Target text	Target image
LEVIR-CD	change mask matching	*Represent the given remote sensing images with the following change detection task.* **Given two remote sensing images of the same location at different times, retrieve the change mask that highlights the differences between them.**		*Represent this change mask for change detection retrieval.*	
SYSU-CD	change mask matching	*Represent the given remote sensing images with the following change detection task.* **Given a pair of remote sensing images from time1 and time2, retrieve the change mask that highlights the differences between them.**		*Represent this change mask for change detection retrieval.*	
VL-CMU-CD	change mask matching	*Represent the given remote sensing images with the following change detection task.* **Given a pair of remote sensing images from time1 and time2, retrieve the change mask that highlights the differences between them.**		*Represent this change mask for change detection retrieval.*	

Table 9 (continued)

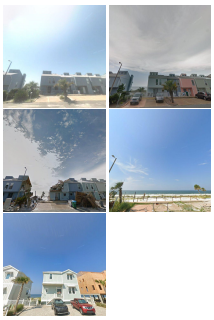
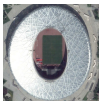
Dataset	Subtask	Query text	Query image	Target text	Target image
OSF Recovery	disaster damage assessment	*Assess the post-disaster recovery status of this location by comparing multi-temporal street-level images.*		*Represent the given post-disaster recovery assessment.* **Improved Reconstruction: The quality of the reconstructed buildings is better than before the disaster.**	-

Table 10: Classification subtask examples.

Dataset	Subtask	Query text	Query image	Target text	Target image
AID	scene classification	*Identify the scene category in the given remote sensing image.*		*Represent the remote sensing scene category.* **stadium**	-
PlacePulse	safety/ liveliness/ beauty/ wealth/ depression/ boringness	*Assess the perceived safety/ liveliness/beauty/wealth/ depression/boringness of this street view image.*		*Represent the safety/ liveliness/beauty/wealth/ depression/boringness perception level of a street view image.* **A street view image with wealth score in [6.5, 7.0)**	-
CityLens	gdp/ population density/ average building height/ healthcare accessibility	*Given a satellite image and multiple street-view images of an urban area, select the normalized GDP/population density/average building height/healthcare accessibility score (0.0 to 10.0) that best describes this area.*		*Represent the given GDP (PPP 2005 international dollars) /population density/average building height/healthcare accessibility score.* **An urban area with health index score in [9.5, 10.0).**	-
UrBench	scene-recognition	*Represent the given image with the following question.* **Which class of object appears in these images? Classes: bus station, fountain, graveyard, memorial or monument, motorway, parking, pitch, rail, residential, stadium, swimming pool, tower. Note that the residential class has the lowest priority and is selected when no other class matches the given image.**		*Represent the user's input.* **parking**	-

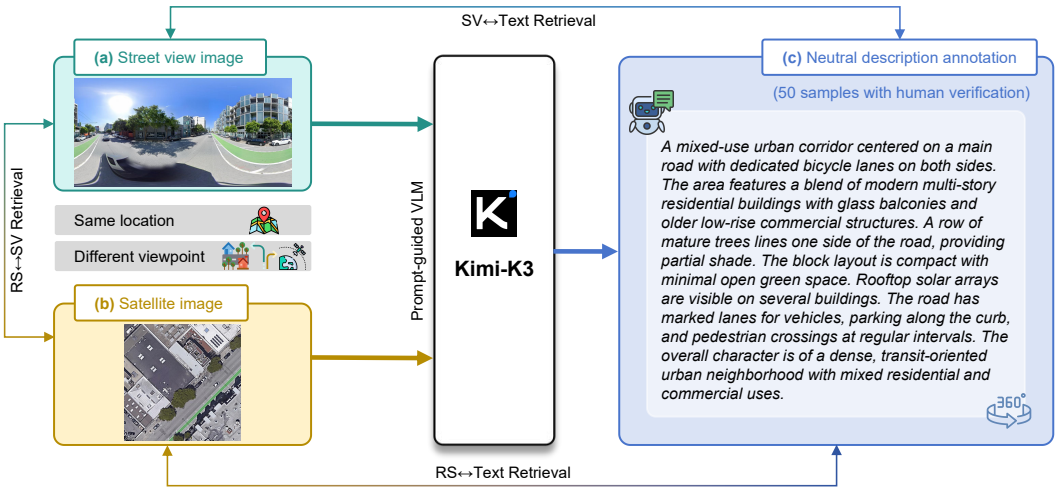


Figure 7: Annotated aligned VIGOR triplets used for instruction-shift analysis.

D Detailed results

Appendix Table 12∼16 report the full task-level results for all 45 GeoMEB evaluation subtasks, grouped by meta-task. These tables complement the aggregate results in Table 2 by showing how each model performs on individual datasets and task formulations, including retrieval directions, VQA subsets, grounding sources, change-detection benchmarks, and classification labels. The detailed breakdown shows that model rankings are not uniform across subtasks. For example, several embedding-specialized models achieve very high scores on Im2GPS3k caption retrieval, but remain weak on GPS-region retrieval from the same dataset. Similarly, change-detection results vary sharply across benchmarks, with some models supporting OSF Recovery but failing to handle mask-based multi-image inputs. This task-level variation supports the need for reporting both aggregate scores and fine-grained results when evaluating geospatial embedding models.

E Geo-Embed Training details

This section summarizes the default training configuration used for Geo-Embed. Geo-Embed is initialized from Qwen3-VL-Embedding-2B and fine-tuned with LoRA-based adaptation under the instruction-conditioned contrastive objective described earlier. Appendix Table 11 reports the main optimization, adaptation, sequence-length, and hardware settings for our final Geo-Embed training run. The LoRA rank and batch size follow the best-performing setting identified in the hyperparameter analysis.

F Baseline model details

This section summarizes the baseline models evaluated in GeoMEB, grouped by model family. We specify the evaluated checkpoint or model variant for each baseline.

CLIP-family models.

- **CLIP** (Radford et al., 2021) is the original contrastive image-text dual encoder trained on large-scale natural image-caption pairs. It represents the standard general-purpose image-text alignment paradigm, where images and text are encoded by separate towers and matched through cosine similarity. We evaluate the CLIP ViT-L/14 checkpoint.
- **OpenCLIP** (Fang et al., 2023) follows the CLIP training framework with open large-scale training data and improved data filtering. It provides a strong open CLIP-family baseline for general image-text retrieval. We evaluate the DFN5B-CLIP-ViT-H-14 checkpoint.

Qualitative example of SV Localization task (i2t/t2i)



A quiet commercial street in China with closed storefronts, parked scooters and bicycles, visible Chinese signage, overcast daylight, eye-level forward-facing view, no pedestrians or moving traffic.

The scene depicts a subdued urban commercial street lined with single-story storefronts beneath weathered multi-story buildings, their facades showing signs of age and exposure. Several shops remain shuttered with metal roll-down doors, while one central store displays an open glass entrance flanked by stacked cardboard boxes and promotional materials. Prominent signage above the shops includes Tricycle, Haiji Tea Shop, Qiang Sheng Trading Company, Jiangnan Pastries, and Big Man, rendered in bold Chinese characters on colored backgrounds—green, orange, black, and red—with some featuring decorative white borders. Parked along the paved sidewalk are two bicycles and a red electric scooter with a rear storage box; further back, a blue three-wheeled utility vehicle is partially visible under an awning. The road surface appears worn, with faint lane markings and scattered debris, suggesting low recent activity. No pedestrians or moving vehicles are present, contributing to a still, almost dormant atmosphere. Lighting is diffuse and even, consistent with overcast daylight, casting soft shadows and muting color saturation. The camera angle is eye-level and forward-facing, capturing the full width of the street frontage without distortion or artistic framing. There are no visible traffic signals, streetlights, or public transit infrastructure. The overall impression is of a typical neighborhood commercial strip during off-hours, possibly early morning or midday lull, where daily commerce has yet to fully activate. This setting aligns with a residential-commercial zone in a smaller Chinese city, where local businesses cater to daily needs and foot traffic remains modest outside peak hours.

Figure 8: A qualitative example of the SV Localization image-to-text and text-to-image tasks. The annotated caption provides a detailed street-level scene description, including storefronts, signage, parked vehicles, lighting conditions, and local commercial context.

Qualitative example of Im2GPS3k caption task (i2t/t2i)



Caption ID: cap_txt_315251374

The scene presents a panoramic view of a densely built European city situated on a steep, terraced hillside that descends to a wide riverbank. The architecture is characterized by tightly packed, multi-story buildings with predominantly red-tiled roofs and facades painted in a variety of muted pastel and earth tones. The river in the foreground is calm, reflecting the sky and the structures along its edge, with several small boats moored or navigating its surface. A broad, paved riverside promenade runs along the water's edge, lined with arches and outdoor seating areas, suggesting commercial or leisure activity. Above the riverfront, the city climbs the hill in successive tiers, with narrow streets and stairways weaving between buildings. Prominent structures include a large, pale, neoclassical-style building with a symmetrical facade and a central dome, situated near the crest of the hill, and a tall, slender church spire visible further back. The sky is clear and blue with minimal cloud cover, indicating fair weather and bright daylight conditions.

The reasoning process for identifying the location is based on several key visual markers:

1. The topography—a steep, densely urbanized hillside descending to a major river—is characteristic of Porto, Portugal, which is built along the Douro River.
2. The architectural style—colorful, tiled-roof buildings clustered tightly on a slope—is emblematic of Porto's Ribeira district.
3. The large, pale neoclassical building near the top of the hill corresponds to the Palácio de Cristal or the adjacent Museu de Arte Contemporânea, located in the Jardim do Morro area overlooking the river.
4. The presence of traditional Rabelo boats, historically used to transport Port wine, and modern tour boats along the riverbank further supports the identification, as this is a common sight in Porto's Ribeira.
5. The overall composition—including the river's curve, the alignment of the buildings, and the vantage point—matches well-known photographic perspectives of Porto from the Dom Luís I Bridge or the opposite bank in Vila Nova de Gaia.

True geographic location: city: Porto; latitude: 41.1496; longitude: -8.6109.

Figure 9: A qualitative example of the Im2GPS3k caption image-to-text and text-to-image tasks. The annotated description provides geolocation-relevant visual and contextual cues, such as topography, riverfront layout, architecture, landmarks, boats, and geographic coordinates, supporting fine-grained location matching.

Table 11: Default training configuration for Geo-Embed.

Item	Setting
Backbone	Qwen3-VL-Embedding 2B
Adaptation	LoRA
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.1
Schedule	Cosine decay
Warmup ratio	0.05
Training epoch	1
Validation split	0.01
LoRA rank / alpha	16 / 32
Temperature	0.02
Batch size	1024
Max sequence length	4096
Precision	bf16
Gradient clipping	1.0
Random seed	42
Gradient Checkpoint	True
Distributed training	MS-Swift + DeepSpeed ZeRO-2
Hardware	4×RTX 4090 + 4×L20 (NVIDIA)

- **RemoteCLIP** (Liu et al., 2024) adapts CLIP-style contrastive learning to remote sensing vision-language data. It is expected to better capture overhead imagery, land-use patterns, and remote-sensing visual semantics than generic image-text encoders. We evaluate the RemoteCLIP-ViT-L-14 checkpoint.
- **GeoRSCLIP** (Zhang et al., 2024) is a geospatial CLIP model trained with RS5M, a large-scale remote-sensing image-text dataset. It provides a domain-specialized baseline for aligning overhead images with natural language descriptions. We evaluate the GeoRSCLIP-ViT-H-14 checkpoint.
- **SkyScript-CLIP** (Wang et al., 2024) is trained on SkyScript, a semantically diverse remote-sensing image-text dataset. It tests whether richer remote-sensing captions improve transfer to heterogeneous urban retrieval and recognition tasks. We evaluate the released SkyScript-CLIP checkpoint.
- **SigLIP2** (Tschannen et al., 2025) is a recent CLIP-family encoder trained with a sigmoid contrastive objective and designed to improve multilingual semantic understanding, localization, and dense visual representations. We evaluate the SigLIP2 SO400M-384 checkpoint.

Instruction-tuned VLMs.

- **GeoChat** (Kuckreja et al., 2024) is a geospatial vision-language model designed for grounded visual instruction following. Although it is not trained as a retrieval embedder, it tests whether geospatial instruction tuning produces transferable embedding representations. We evaluate the GeoChat-7B checkpoint.
- **Qwen3-VL-Instruct** (Bai et al., 2025) is a general-purpose instruction-tuned vision-language model from the Qwen3-VL family. It tests whether multimodal instruction-following ability can be directly reused for retrieval in GeoMEB. We evaluate Qwen3-VL-Instruct-2B and Qwen3-VL-Instruct-8B.

Embedding-specialized VLMs.

- **UME-R1** (Lan et al., 2025) explores reasoning-driven generative multimodal embeddings. It represents a recent direction that uses reasoning-oriented training to improve embedding quality beyond conventional image-text contrastive alignment. We evaluate the UME-R1-7B checkpoint.
- **VLM2Vec-V2.0** (Meng et al., 2025) extends VLM2Vec-style training to broader multimodal embedding settings, including images, videos, and visual documents. It serves as

a strong open embedding-specialized VLM baseline. We evaluate the VLM2Vec-V2.0-3B checkpoint.

- **OmniEmbed-Nemotron** (Xu et al., 2025) is a unified multimodal retrieval model covering text, image, audio, and video. Its training objective emphasizes cross-modal retrieval across multiple input types, making it a relevant baseline for testing whether general omni-modal embedding transfers to geospatial tasks. We evaluate the OmniEmbed-Nemotron-3B checkpoint.
- **Qwen3-VL-Embedding** (Li et al., 2026a) is a native multimodal embedding model built on the Qwen3-VL family and one of the strongest open embedding-specialized VLMs in our comparison. It provides strong pretrained cross-modal alignment for general-purpose retrieval, making it both a competitive baseline and the initialization backbone for Geo-Embed. We evaluate Qwen3-VL-Embedding-2B and Qwen3-VL-Embedding-8B.
- **Seed1.6-Embedding** (ByteDance Seed Team, 2025) is a closed-source multimodal embedding model released as part of the Seed1.6 model family. It targets general-purpose multimodal retrieval and serves as one of the strongest proprietary baselines in our experiments. We evaluate it through the `doubao-embedding-vision-251215` API.

Table 12: Detailed results on retrieval tasks.

Dataset	Subtask	Metric	CLIP	OpenCLIP	RemoteCLIP	GeoRSCLIP	SkyScript-CLIP	SigLIP2	GeoChat	Qwen3-VL-Instruct 2B	Qwen3-VL-Instruct 8B	UME-R1	VLM2Vec-V2.0	OmniEmbed-Nemotron	Seed1.6-Embedding	Qwen3-VL-Embedding 8B	Qwen3-VL-Embedding 2B	Geo-Embed
RSICD	i2t	Mean R@1/5/10	18.88	27.08	36.47	30.83	19.82	27.69	0.46	0.37	1.04	26.96	26.72	14.91	25.37	30.59	31.93	36.29
RSICD	t2i	Mean R@1/5/10	18.79	28.53	35.28	29.65	21.03	27.52	0.48	0.72	0.82	27.70	28.58	15.53	26.10	34.01	30.57	33.40
UAV-GeoLoc	country	R@1	7.70	18.09	0.17	15.62	12.85	9.93	0.46	0.00	0.00	8.13	7.39	4.50	7.67	13.52	13.70	29.20
UAV-GeoLoc	terrain	R@1	11.33	18.31	0.75	22.73	19.23	5.62	0.87	0.02	0.16	8.57	10.28	6.28	11.41	17.00	17.74	34.41
VIGOR	–	R@1	1.49	3.42	0.26	1.75	1.75	2.41	0.07	0.00	0.22	2.38	3.16	2.60	2.86	2.27	1.78	3.12
SV Localization	i2t	Mean R@1/5/10	7.40	23.27	5.12	18.21	7.47	15.95	0.01	0.03	0.03	1.30	81.61	38.72	82.47	97.76	92.93	94.97
SV Localization	t2i	Mean R@1/5/10	9.11	26.16	4.39	20.88	10.69	14.92	0.01	0.03	0.03	13.54	84.40	17.36	85.30	96.43	92.48	95.82
Im2GPS3k	GPS i2t	R@1	0.07	0.03	0.03	0.07	0.03	0.07	0.37	0.03	0.53	5.37	0.17	0.90	9.54	8.48	0.83	2.00
Im2GPS3k	GPS t2i	R@1	0.00	0.09	0.00	0.09	0.00	0.09	0.09	0.09	0.00	0.44	0.18	0.09	2.63	2.54	0.26	0.26
Im2GPS3k	Caption i2t	Mean R@1/5/10	84.41	94.35	65.96	86.81	85.10	75.36	2.90	0.37	0.74	62.92	94.16	96.12	98.95	99.10	99.48	99.74
Im2GPS3k	Caption t2i	Mean R@1/5/10	82.53	93.28	52.23	83.23	84.65	76.40	2.88	0.68	1.37	90.49	97.64	94.29	99.37	99.65	99.34	99.76
VRSBench	caption i2t	Mean R@1/5/10	7.02	14.01	4.94	16.14	8.68	7.89	0.58	0.11	0.59	11.08	15.34	8.36	18.12	23.32	16.63	26.52
VRSBench	caption t2i	Mean R@1/5/10	6.36	13.84	3.65	14.90	8.34	9.81	0.27	0.09	0.12	14.92	17.90	5.02	17.93	25.66	18.67	25.12
Average			**19.62**	**27.73**	**16.10**	**26.22**	**21.51**	**21.05**	**0.73**	**0.20**	**0.43**	**21.06**	**35.96**	**23.44**	**37.52**	**42.33**	**39.72**	**44.66**

Table 13: Detailed results on visual question answering tasks.

Dataset	Subtask	Metric	CLIP	OpenCLIP	RemoteCLIP	GeoRSCLIP	SkyScript-CLIP	SigLIP2	GeoChat	Qwen3-VL-Instruct 2B	Qwen3-VL-Instruct 8B	UME-R1	VLM2Vec-V2.0	OmniEmbed-Nemotron	Seed1.6-Embedding	Qwen3-VL-Embedding 8B	Qwen3-VL-Embedding 2B	Geo-Embed
VRSBench	vqa	R@1	0.99	0.93	1.48	0.41	0.89	0.09	0.02	0.04	0.09	35.45	0.52	0.06	14.33	40.05	34.42	35.66
MME-RealWorld	AutonomousDriving	R@1	18.10	22.43	23.67	21.04	18.56	18.10	18.72	21.81	22.04	33.87	28.92	17.56	34.57	37.97	36.27	42.23
MME-RealWorld	MME-HD-CN	R@1	2.19	9.50	9.06	14.77	2.49	3.80	15.64	28.22	27.19	55.85	36.55	28.22	61.40	68.27	64.47	59.36
MME-RealWorld	diagram_and_table	R@1	4.80	6.86	3.77	4.37	4.46	0.26	10.37	24.68	32.56	64.87	32.73	9.68	55.01	76.18	66.67	60.41
MME-RealWorld	monitoring_images	R@1	11.23	11.68	11.68	10.18	10.63	9.43	18.41	14.97	16.62	39.37	15.42	10.18	26.50	41.77	38.77	39.82
MME-RealWorld	ocr_cc	R@1	6.72	20.15	7.36	7.12	6.72	2.75	9.55	29.61	32.20	61.08	40.94	21.44	60.92	72.25	67.15	62.94
MME-RealWorld	remote_sensing	R@1	7.58	10.26	8.74	8.39	7.69	9.32	10.61	26.11	20.05	47.32	20.75	11.89	43.12	56.88	47.79	49.30
UrBench	counting	R@1	20.45	27.27	7.95	18.18	19.32	20.45	27.27	1.14	15.91	22.73	2.27	11.36	43.18	44.32	23.86	39.77
UrBench	object-attribute-recognition oe	R@1	3.53	0.00	1.18	3.53	3.53	25.88	2.35	11.76	10.59	28.24	14.12	2.35	15.29	4.71	4.71	27.06
UrBench	road-understanding oe	R@1	50.98	49.02	49.02	50.98	50.98	50.98	50.98	49.02	49.02	50.98	47.06	50.98	50.98	56.86	50.98	84.31
UrBench	role-based-reasoning	R@1	1.78	4.00	0.89	1.78	2.22	1.78	0.44	0.00	0.89	22.67	6.22	0.44	19.56	41.78	33.33	28.44
UrBench	traffic-sign-reasoning	R@1	1.00	1.00	2.00	2.00	2.00	2.00	1.00	2.00	2.00	11.00	2.00	1.00	12.00	17.00	14.00	11.00
UrBench	visual-prompt-reasoning	R@1	2.00	2.00	0.00	0.00	2.00	0.00	4.00	2.00	4.00	20.00	8.00	4.00	24.00	30.00	28.00	26.00
Average				**10.10**	**12.70**	**9.75**	**10.98**	**10.11**	**11.14**	**13.03**	**16.26**	**17.94**	**37.96**	**19.65**	**13.01**	**35.45**	**45.23**	**39.26**	**43.56**

Table 14: Detailed results on grounding tasks.

Dataset	Subtask	Metric	CLIP	OpenCLIP	RemoteCLIP	GeoRSCLIP	SkyScript-CLIP	SigLIP2	GeoChat	Qwen3-VL-Instruct 2B	Qwen3-VL-Instruct 8B	UME-R1	VLM2Vec-V2.0	OmniEmbed-Nemotron	Seed1.6-Embedding	Qwen3-VL-Embedding 8B	Qwen3-VL-Embedding 2B	Geo-Embed
Cityscapes	–	R@1	1.72	1.84	1.20	2.20	1.56	2.56	0.04	0.04	0.08	6.08	4.28	2.00	8.56	8.16	8.20	10.20
Mapillary	–	R@1	0.78	1.47	0.28	0.81	0.76	2.32	0.01	0.03	0.02	3.03	2.22	1.92	4.19	5.87	1.79	8.06
VRSBench	referring	R@1	4.16	5.41	0.95	5.36	5.58	5.08	0.53	0.01	0.59	4.17	5.85	2.81	7.68	6.27	3.56	13.00
Average			**2.22**	**2.91**	**0.81**	**2.79**	**2.63**	**3.32**	**0.19**	**0.03**	**0.23**	**4.43**	**4.12**	**2.24**	**6.81**	**6.77**	**4.52**	**10.42**

Table 15: Detailed results on change detection tasks. For models that do not support multi-image inputs, "–" denotes unavailable results and is counted as 0 when computing averages.

Dataset	Subtask	Metric	CLIP	OpenCLIP	RemoteCLIP	GeoRSCLIP	SkyScript-CLIP	SigLIP2	GeoChat	Qwen3-VL-Instruct 2B	Qwen3-VL-Instruct 8B	UME-R1	VLM2Vec-V2.0	OmniEmbed-Nemotron	Seed1.6-Embedding	Qwen3-VL-Embedding 8B	Qwen3-VL-Embedding 2B	Geo-Embed
LEVIR-CD	–	R@1	2.34	0.78	1.56	1.56	3.12	0.78	0.78	0.00	0.78	0.00	0.78	0.78	3.12	2.34	1.56	5.47
SYSU-CD	–	R@1	-	-	-	-	-	-	-	0.00	0.03	0.10	0.07	0.00	0.20	0.07	0.10	0.83
VL-CMU-CD	–	R@1	-	-	-	-	-	-	-	0.23	0.23	0.47	0.70	1.17	2.10	0.93	0.23	2.10
OSF Recovery	–	R@1	-	-	-	-	-	-	-	41.61	33.58	37.96	29.93	34.31	42.34	31.68	35.77	37.96
Average			**0.58**	**0.20**	**0.39**	**0.39**	**0.78**	**0.20**	**0.20**	**10.46**	**8.65**	**9.63**	**7.87**	**9.07**	**11.94**	**8.75**	**9.42**	**11.59**

Table 16: Detailed results on classification tasks.

Dataset	Subtask	Metric	CLIP	OpenCLIP	RemoteCLIP	GeoRSCLIP	SkyScript-CLIP	SigLIP2	GeoChat	Qwen3-VL-Instruct 2B	Qwen3-VL-Instruct 8B	UME-R1	VLM2Vec-V2.0	OmniEmbed-Nemotron	Seed1.6-Embedding	Qwen3-VL-Embedding 8B	Qwen3-VL-Embedding 2B	Geo-Embed
AID	scene classification	Acc.	68.80	74.65	81.75	74.95	72.40	67.60	27.75	2.75	2.75	62.80	63.50	46.30	71.30	70.75	66.20	76.00
PlacePulse	safety	R@1	1.60	3.85	0.29	10.53	1.27	2.07	12.38	3.70	7.66	11.62	4.28	8.53	12.45	4.94	2.80	9.01
PlacePulse	lively	R@1	0.31	6.35	0.81	11.13	0.17	6.65	7.36	0.64	0.10	0.68	6.79	5.02	6.41	0.14	4.51	3.95
PlacePulse	beautiful	R@1	5.13	5.33	1.05	7.70	4.67	10.59	13.42	1.58	0.33	6.78	8.03	10.13	11.64	1.25	5.46	4.92
PlacePulse	wealthy	R@1	1.90	0.15	0.29	8.45	1.17	1.17	10.28	1.31	0.07	5.32	8.09	9.33	11.81	0.58	13.70	7.58
PlacePulse	depressing	R@1	4.20	1.78	0.10	5.98	3.46	7.87	11.96	5.77	0.00	6.30	2.94	6.72	0.52	0.00	6.61	6.40
PlacePulse	boring	R@1	3.52	1.99	0.15	7.03	5.66	3.67	11.16	3.06	0.00	6.12	6.57	5.05	3.98	0.00	3.52	5.05
CityLens	gdp	R@1	5.10	4.30	4.90	4.20	6.40	4.90	6.30	6.90	6.20	5.80	5.90	5.90	4.40	9.30	6.20	8.50
CityLens	pop	R@1	5.00	5.80	5.20	6.00	5.60	3.90	4.80	9.00	5.80	5.30	9.20	7.60	6.00	9.50	6.80	10.30
CityLens	height	R@1	4.40	4.40	4.00	6.20	4.50	4.60	6.30	5.40	5.20	5.40	5.40	5.40	7.60	8.80	5.40	8.50
CityLens	health	R@1	12.40	14.90	0.40	2.90	15.00	15.50	5.20	0.00	24.20	3.80	3.80	0.60	7.70	5.40	4.60	8.30
UrBench	scene-recognition	R@1	26.67	9.17	33.33	12.50	21.67	5.00	3.33	19.17	10.00	39.17	37.50	21.67	41.67	37.50	55.00	57.50
Average			**11.58**	**11.06**	**11.02**	**13.13**	**11.83**	**11.13**	**10.02**	**4.94**	**5.19**	**13.25**	**13.50**	**11.02**	**15.46**	**12.35**	**15.07**	**17.17**