WIDEUAV: CROSS-VIEW SATELLITE-TILE RETRIEVAL AND TARGET-EXTENT GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

In wide-area satellite imagery, finding the matching tile is only the first step toward localizing a target observed from a UAV: the target must also be grounded within that tile. We study *cross-view geo-grounding*, a unified setting that combines satellite-tile retrieval with target grounding. To support this task, We introduce **WideUAV**, a rendered benchmark comprising 3,505 wide satellite tiles annotated with off-centered target bounding boxes and 112,160 UAV queries captured under controlled variations in altitude and heading. In contrast to existing cross-view benchmarks such as University-1652, where satellite images are tightly cropped to $120\mathrm{m} \times 120\mathrm{m}$ around the target, WideUAV uses $2370\mathrm{m} \times 1328\mathrm{m}$ satellite images, covering over $200\times$ the area. The drone-view queries are sampled at 32 views per location on a fixed grid of 4 altitudes and 8 headings. We further propose UNISAG (**Uni**fied **S**atellite Retrieval **a**nd **G**rounding), a single model that shares visual representations across retrieval and grounding while incorporating UAV heading and altitude cues. To jointly evaluate retrieval and grounding, we introduce retrieval-gated uIoU, an end-to-end metric that assigns zero uIoU when the correct satellite tile is not retrieved. On this metric, UNISAG achieves 22.67%, outperforming several competitive re-implemented approaches, including TransGeo, with an 18.87-point gain over the strongest baseline at 3.80%. We also observe an asymmetric interaction between the two parts: grounding supervision improves the retrieval representation, whereas retrieval supervision improves little to grounding.

1 INTRODUCTION

Visual geo-localization provides a complementary cue for UAV navigation when reliable positioning signals are unavailable or degraded. Existing drone-to-satellite benchmarks largely formulate cross-view localization as image retrieval, where a UAV image is matched against geo-referenced satellite candidates that are typically cropped around the target (Zheng et al., 2020; Zhu et al., 2022; 2023). This formulation is effective for cross-view matching, but largely removes the spatial search required over wide-area satellite imagery. In such a setting, finding the matching tile is only the first step: the target observed from the UAV must also be localized within that tile. We therefore study *cross-view geo-grounding*, a unified setting that combines satellite-tile retrieval with target grounding (Figure 1), where retrieval determines *which satellite region to search* and grounding determines *where the target lies within that region*. Existing benchmarks capture only parts of this problem: retrieval datasets typically use compact target-centered satellite crops, whereas grounding benchmarks assume that the relevant satellite region is already given (Sun et al., 2023b; Huang et al., 2025). As summarized in Table 1, few benchmarks jointly require wide-area satellite search and precise target grounding.

To support this setting, we introduce **WideUAV**, a rendered benchmark comprising 3,505 wide satellite tiles annotated with target bounding boxes and 112,160 UAV queries captured under controlled variations in altitude and heading. Compared with University-1652, whose satellite images cover approximately $120\,\mathrm{m} \times 120\,\mathrm{m}$ around the target, WideUAV uses raw satellite regions of $2370\,\mathrm{m} \times 1328\,\mathrm{m}$, covering over $200\times$ the area. Targets are not constrained to the tile center, and each location is observed from 32 UAV viewpoints spanning four altitudes and eight headings. WideUAV therefore preserves the surrounding geographic context needed for cross-view matching while introducing substantial spatial uncertainty within each retrieved satellite region.

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.

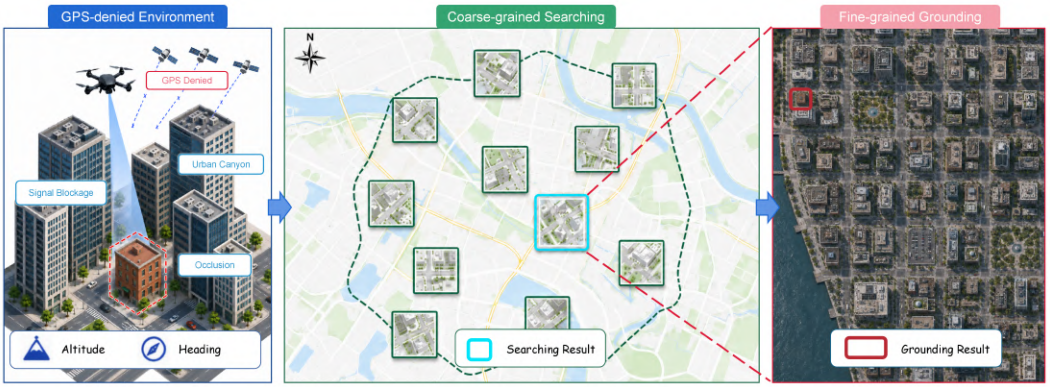

Figure 1: **Cross-view geo-grounding.** A UAV query is localized by first retrieving a satellite tile and then predicting the target extent within it. The wide tile preserves surrounding context while requiring the model to distinguish the target from nearby buildings. Heading and altitude provide geometric context for grounding.

Table 1: **Cross-view localization benchmarks.** The datasets are coarsely classified as retrieval (*top*) and grounding dataset (*bottom*). Among these datasets, DenseUAV (Dai et al., 2023) studies both retrieval and grounding, but its footprint and resolution are relatively limited.

Benchmark	Scope	BBox/Point	UAV View	Max Footprint (m)	Resolution	# UAV	# Sat.
University-1652 (Zheng et al., 2020)	72 univ.	×	Oblique	120×120	512×512	89,208	1,652
SUES-200 (Zhu et al., 2023)	1 univ.	×	Oblique	120×120	512×512	16,000	200
CVOGL (Sun et al., 2023b)	U.S.	✓	Nadir	1024×1024	1024×1024	5,279	**5,836**
DenseUAV (Dai et al., 2023)	14 univ.	✓	Nadir	41×41	512×512	2,331	3,033
UAV-VisLoc (Xu et al., 2024)	8 cities	×	Nadir	$\sim$ **6000²**	$\sim$ **20,000²**	6,742	11
AnyVisLoc (Ye et al., 2026)	15 cities	×	Oblique	$\sim 3000^2$	$\sim 10{,}000^2$	18,000	25
WideUAV	**206 univ. & 13 cities**	✓	**Oblique**	2370×1328	3840×2160	**112,160**	3,505

Max Footprint and Resolution refer to raw satellite imagery. WideUAV test tiles cover approximately $1183\ \mathrm{m} \times 660\ \mathrm{m}$ and are resized to 768×432 pixels. BBox denotes target-extent annotations. Footprints are estimates, and $\sim$ marks rounded values.

The wide-area setting also changes how retrieval and grounding should interact. Directly grounding over all satellite regions is computationally expensive, while aggressively reducing the satellite resolution removes fine-grained structures needed for precise localization. This naturally suggests a coarse-to-fine formulation: first retrieve a context-preserving satellite region, then ground the target within it. We introduce UNISAG (**Uni**fied **S**atellite Retrieval **a**nd **G**rounding), a single model that shares visual representations across both components. Regional soft labels supervise retrieval using the satellite region containing the target center, while heading-conditioned correlation reuses dense satellite features to predict the target bounding box conditioned on the UAV viewpoint.

To evaluate the complete localization pipeline, we further introduce *retrieval-gated uIoU*, which jointly measures retrieval and grounding by assigning zero uIoU when the correct satellite tile is not retrieved. UNISAG achieves 22.67% retrieval-gated uIoU, outperforming several competitive re-implemented approaches by 18.87 points over the strongest baseline. Beyond the overall performance gain, our experiments also reveal an asymmetric interaction between retrieval and grounding: grounding supervision improves the retrieval representation, whereas retrieval supervision offers little benefit to grounding. Our main contributions are summarized as follows:

- **A benchmark for cross-view geo-grounding.** WideUAV combines wide satellite tiles, off-center test targets, calculated bboxes, and controlled acquisition geometry under a target-ID-disjoint split.
- **A unified retrieval-and-ground model.** UNISAG connects region-level supervision for tile retrieval and heading-conditioned grounding through a shared visual representation.
- **An end-to-end evaluation and empirical study.** We introduce retrieval-gated uIoU for evaluating the complete localization pipeline and show that UNISAG substantially outperforms competitive baselines, while revealing an asymmetric interaction between grounding and retrieval supervision.

2 Related Work

Cross-view geo-localization and UAV benchmarks. CVM-Net (Hu et al., 2018) learns global descriptors for ground-to-aerial matching and TransGeo (Zhu et al., 2022) models correspondence with transformers and attention-guided cropping. Sample4Geo (Deuser et al., 2023) improves contrastive learning with geographically and visually informed hard negatives. VIGOR (Zhu et al., 2021) supports off-center queries and metric localization, while Ret4Loc (Kalantidis et al., 2024) combines synthetic appearance variation and geometric consistency for visual-localization retrieval. UAV benchmarks cover worldwide university buildings (University-1652 (Zheng et al., 2020)), flight-height variation (SUES-200 (Zhu et al., 2023)), and real UAV imagery paired with high-resolution satellite tiles (UAV-VisLoc (Xu et al., 2024)). OrthoLoC (Dhaouadi et al., 2026) supports pose estimation with orthographic geodata. WideUAV extends the University-1652 dataset with wide satellite context and off-center targets for joint evaluation of gallery retrieval and target grounding.

Dense representations and geometric correspondence. Image-level pretraining does not directly ensure accurate local correspondence. CLIPSelf (Wu et al., 2024) aligns regional features with representations of corresponding image crops, while spatial correlation distillation (Qiu et al., 2025) preserves and refines CLIP spatial structure. Register tokens (Darcet et al., 2024) reduce transformer feature-map artifacts and NeCo (Pariza et al., 2025) improves dense representations through consistent patch-neighbor rankings. Win-Win (Leroy et al., 2024) addresses high-resolution dense-prediction costs through training on selected windows. For cross-view correspondence, GIM (Shen et al., 2024) learns generalizable matching through self-training on internet videos, and multiview equivariance finetuning (You et al., 2025) improves cross-view feature consistency and usefulness for geometric tasks. These studies motivate local feature quality for a shared encoder supporting tile ranking and spatial localization.

Cross-view object grounding and unified localization. Deformable DETR (Zhu et al., 2020) and DINO (Zhang et al., 2022) improve spatial attention and box prediction. To associate a query with its target across views, DetGeo (Sun et al., 2023b) incorporates query-aware fusion, OCGNet (Huang et al., 2025) introduces location cues and cross-attention, and TROGeo (Zhang & Zhu, 2025) predicts segmentation masks. UnifyGeo (Song et al., 2025) combines retrieval and metric localization through shared representations. Our UniSAG couples region-level tile-retrieval supervision with dense cross-view matching and grounding conditioned on known flight geometry through a shared encoder. We measure the resulting trade-off with task specialization on WideUAV.

3 WideUAV

Dataset overview. WideUAV is a rendered benchmark for cross-view geo-grounding, containing 3,505 satellite tiles and 112,160 UAV queries across 206 universities and 13 city scenes. Each target location is associated with one satellite tile annotated with a target bounding box and 32 UAV queries captured from four altitudes (150, 200, 250, and 300 m) and eight headings spaced by 45°. All UAV views use a fixed 45° pitch, providing controlled variation in altitude and heading while retaining oblique viewpoints representative of forward-looking flight. WideUAV spans 6 continents, 33 countries, and 222 city labels (Figure 2A), covering diverse urban layouts and visual environments. Raw satellite imagery covers up to $2370\,\mathrm{m} \times 1328\,\mathrm{m}$, while evaluation tiles cover approximately $1183\,\mathrm{m} \times 660\,\mathrm{m}$ and are resized to 768×432 pixels for model input. The evaluation footprint is about $54\times$ the area of the $120\,\mathrm{m} \times 120\,\mathrm{m}$ satellite images in University-1652 and over $400\times$ that of the $41\,\mathrm{m} \times 41\,\mathrm{m}$ tiles in DenseUAV (Figure 2B and Table 1).

Unlike target-centered retrieval benchmarks, WideUAV distributes targets across the interior of the satellite tiles (Figure 2D), with mean and median displacements of 299.94 m and 291.78 m from the tile center. The wide satellite context also introduces visually similar distractors: neighboring tiles can share recurring rooftops, roads, and urban layouts despite corresponding to different target locations (Figure 2E). Figure 2C further illustrates the visual diversity of UAV observations through a t-SNE projection of SigLIP2 features under a fixed altitude and heading. These properties make WideUAV substantially different from conventional cross-view retrieval benchmarks: the satellite tile provides broad geographic context but does not directly reveal the target position, requiring both satellite-region retrieval and within-tile grounding. Details of the rendering pipeline, camera configuration, target-box generation, and preprocessing are provided in Appendix A.

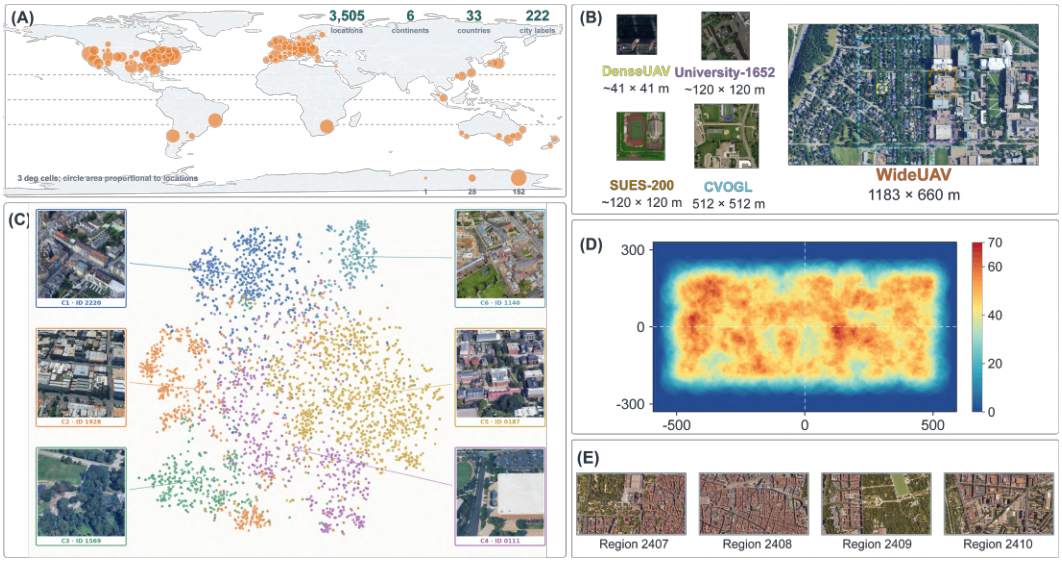

Figure 2: **Geographic and visual diversity of WideUAV.** (A) Geographic distribution of the 3,505 target locations across six continents, 33 countries, and 222 city labels. (B) Satellite footprints compared with representative cross-view benchmarks. (C) t-SNE visualization of SigLIP2 query features at a fixed altitude and heading, with representative samples from six unsupervised visual groups. (D) Distribution of target centers within the satellite-tile frame. (E) Neighboring Madrid tiles illustrating visually similar urban distractors.

Table 2: **WideUAV split statistics.** Each satellite tile is paired with 32 drone queries (4 altitudes $\times$ 8 headings). All views of each target location belong to a single split.

Split	# Sat. tiles	# Drone queries
Training	701	22,432
Test	2,804	89,728
Total	**3,505**	**112,160**

Table 3: **One benchmark, three questions.** We evaluate retrieval-gated geo-grounding alongside tile retrieval and oracle grounding.

Level	Input	Output / metric
Tile retrieval	query / sat. gallery	tile rank / R@K
Oracle grounding	query / GT tile	box / mIoU, Acc, CDE
Retrieval-gated diagnostic	query / sat. gallery	box / uIoU, uAcc, uCDE

Dataset split. We partition WideUAV by target location, assigning all 32 UAV views of the same target to a single split. The resulting training and test sets contain 701 and 2,804 target locations, corresponding to 22,432 and 89,728 UAV queries, respectively (Table 2). The split is target-ID-disjoint rather than strictly campus-disjoint, providing a large held-out satellite gallery for evaluating retrieval and grounding on unseen targets. Geographic proximity between training and test locations is analyzed in Appendix B.3.

4 UniSAG: Unified Satellite Retrieval and Grounding

WideUAV casts cross-view localization as retrieval followed by within-tile grounding. We therefore design UniSAG to share visual representations between these two components while using flight geometry to guide fine-grained grounding. Figure 3 gives an overview of the framework.

Problem formulation. Given a UAV query image I_d, its heading θ measured clockwise from geographic north, its altitude a, and a satellite gallery $\mathcal{S} = \{I_s^{(j)}\}_{j=1}^{M}$, the goal is to retrieve the satellite tile containing the designated target and predict its axis-aligned bounding box within that tile. We denote the ground-truth tile index by j^* and its target box by b^*. The model therefore predicts $(\hat{j}, \hat{b}) = f(I_d, \theta, a, \mathcal{S})$, where $\hat{j}$ identifies the retrieved tile and $\hat{b}$ localizes the target within it.

Shared encoding and tile retrieval. A shared visual encoder E_v extracts dense UAV and satellite features $\mathbf{F}_d$ and $\mathbf{F}_s$, followed by pooling function as the global drone descriptor $\mathbf{z}_d$ and regional

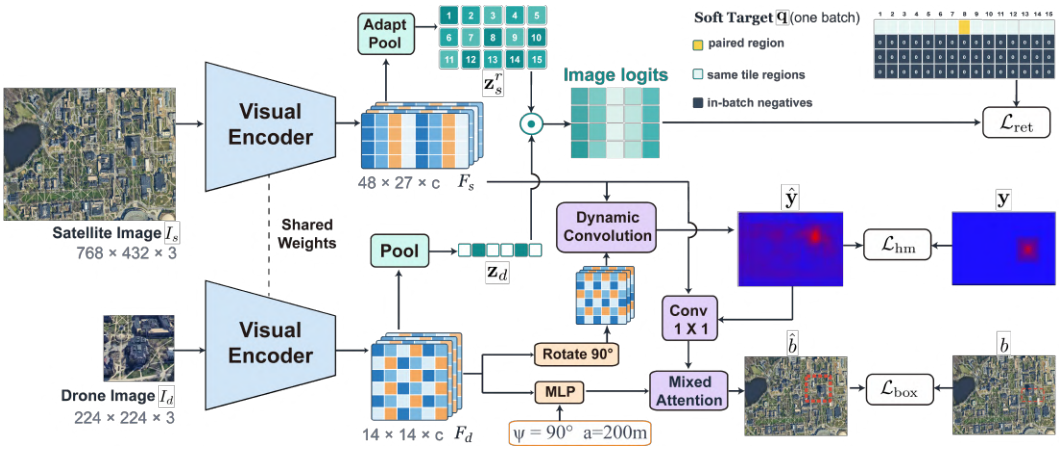

Figure 3: **Overview of UniSAG.** A shared visual encoder supplies pooled descriptors for tile retrieval and dense features for grounding. We align the drone features according to heading for correlation with the satellite map; the response and flight geometry guide Transformer fusion and box prediction. The Transformer module is labeled Mixed Attention in the diagram. The pooled satellite and drone descriptors, soft target, and rotation angle in the diagram correspond to $\mathbf{Z}_s$, $\mathbf{z}_d$, $\mathbf{q}$, and θ in the text.

satellite descriptors $\mathbf{Z}_s$, respectively:

$$\begin{cases} \mathbf{F}_d = E_v(I_d), & \mathbf{z}_d = \mathrm{Pool}(\mathbf{F}_d), \\ \mathbf{F}_s = E_v(I_s), & \mathbf{Z}_s = \mathrm{Pool}(\mathbf{F}_s). \end{cases} \tag{1}$$

Pooling aggregates the full drone map and each satellite region. After ℓ_2 normalization, the maximum regional similarity between $\mathbf{Z}_s$ and $\mathbf{z}_d$ determines the tile ranking.

Regional soft-label retrieval. We identify the positive satellite region by the center of the target box b. As in label smoothing (Szegedy et al., 2016), the soft target $\mathbf{q}$ mixes its one-hot label with a uniform distribution over the regions of the paired tile, using weights $1-\epsilon$ and ϵ. Other tiles receive zero mass. The retrieval loss is

$$\mathcal{L}_{\text{ret}} = -\mathbb{E}\left[\mathbf{q}\,\log \mathrm{softmax}\left(\frac{\mathbf{z}_d \cdot \mathbf{Z}_s}{\tau}\right)\right]. \tag{2}$$

Here τ is the temperature. During training, $\mathbf{Z}_s$ collects all satellite-region descriptors in the batch; the softmax and cross-entropy reduction span those descriptors.

Geometry-conditioned grounding and supervision. Heading θ and altitude a condition the drone features through FiLM (Perez et al., 2018). In the spatial matching branch, we align the drone features according to heading and correlate them with the satellite features:

$$\hat{\mathbf{y}} = \sigma\big(\mathrm{Conv}\big(\mathbf{F}_s, \mathbf{F}_d \circ \mathbf{R}_\theta\big)\big), \qquad \mathbf{R}_\theta = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The rotation acts on the spatial coordinates of $\mathbf{F}_d$. Conv denotes correlation with the aligned drone features, and σ is the sigmoid. The target heatmap $\mathbf{y}$ is constructed from b using a box-size-normalized Chebyshev-distance map: it peaks at the center, decays toward the boundary, and is zero outside the box. We supervise $\hat{\mathbf{y}}$ against $\mathbf{y}$ with focal loss (Lin et al., 2017), denoted by $\mathcal{L}_{\text{hm}}$.

The correlation response is fused with satellite features before a Transformer combines the two views and flight geometry. The predicted heatmap also enhances box confidence through a detached path. A detached confidence prior $\hat{\mathbf{y}}'$ derived from the predicted heatmap $\hat{\mathbf{y}}$ is added to the learned confidence logits:

$$\hat{\mathbf{e}} = \mathrm{Transformer}(\mathbf{F}_s, \mathbf{F}_d, a, \theta, \mathrm{pos_embedding}) + \hat{\mathbf{y}}'.$$

Here Transformer includes the fusion and confidence prediction, and pos_embedding denotes learned spatial embeddings. The YOLO-style head (Redmon & Farhadi, 2018) also predicts the box $\hat{b}$. With one-hot confidence target $\mathbf{e}$, the box objective is

$$\mathcal{L}_{\text{box}} = \mathbb{E}\left[\|\hat{b} - b\|_2^2 - \mathbf{e}\,\log \mathrm{softmax}(\hat{\mathbf{e}})\right]. \tag{3}$$

Table 4: **Tile retrieval on WideUAV.** Each of the 89,728 queries is evaluated against 100 candidate tiles sampled from 2,804 test tiles. We train UNISAG-R for retrieval only and the full model jointly.

Method	R@1 (%)↑	R@5 (%)↑	R@10 (%)↑
Finetuned baselines			
EVA-CLIP (Sun et al., 2023a)	21.88	50.49	64.39
Sample4Geo (Deuser et al., 2023)	27.07	57.58	71.84
SigLIP2 (Tschannen et al., 2025)	33.28	65.23	77.52
CLIP (Radford et al., 2021)	43.02	74.51	83.93
OpenCLIP (Ilharco et al., 2021)	44.48	76.08	84.99
UNISAG-R	62.03	87.21	93.10
UNISAG, full	**65.75**	**90.57**	**95.16**

The L2 term compares boxes in the regression parameterization of the detector. At inference, the highest-confidence candidate supplies the box.

Joint training. Retrieval and paired-tile grounding share the visual encoder and are optimized together:

$$\mathcal{L}_{\text{total}} = (1 - \lambda_{\text{box}})\mathcal{L}_{\text{ret}} + \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{hm}}\mathcal{L}_{\text{hm}}. \tag{4}$$

The loss weights balance the two tasks and heatmap supervision.

5 EXPERIMENT

Evaluation protocol. We evaluate cross-view geo-grounding at three complementary levels (Table 3), separating the two components while also measuring the complete localization pipeline. First, *tile retrieval* evaluates whether the correct satellite region can be identified from the gallery, using Recall@K. Second, *oracle grounding* supplies the ground-truth satellite tile and evaluates within-tile localization using mIoU, Acc@t, and center-distance error (CDE). Third, *retrieval-gated grounding* evaluates the full cross-view geo-grounding setting, where grounding performance is conditioned on successful top-1 retrieval. For retrieval-gated grounding, we introduce **uIoU** as the primary metric: the oracle grounding IoU is retained when the top-1 retrieved tile is accepted and set to zero otherwise. We further extend this protocol with uAcc@t, which measures the fraction of all queries whose gated IoU exceeds threshold t, and uCDE, which reports center-distance error for queries with accepted top-1 retrieval. Detailed definitions of accepted-tile matching, coordinate conversion, and metric implementation are provided in Appendix B.

Implementation details. We evaluate UNISAG and the adapted baselines on the WideUAV test split (Table 2). For UNISAG, we keep the pretrained SigLIP2-B/16 backbone (Tschannen et al., 2025) frozen and train LoRA adapters (Hu et al., 2021) of rank 8, scaling 16, and dropout 0.05 alongside the task-specific modules. Drone inputs are resized to 224×224, while satellite training crops and test tiles are resized to 768×432. At evaluation, we retain the full test tile unless an extent crop is specified. We train for 20 epochs with AdamW, an initial learning rate of 5×10^{-5}, and weight decay 0.01. The batch size is 32. We apply epoch-wise cosine annealing (Loshchilov & Hutter, 2016) over a single cycle, and use FP16 mixed precision with gradient-norm clipping at 1.0. We evaluate the final-epoch checkpoint because of the two task targets. For the full model, we set $\lambda_{\text{box}} = 0.5$ and $\lambda_{\text{hm}} = 0.01$ in Equation 4. Feature geometry, grounding implementation, and detailed loss settings are provided in Appendix F.

Baselines. We train the retrieval baselines on the WideUAV training split: CLIP (Radford et al., 2021), OpenCLIP (Ilharco et al., 2021), EVA-CLIP (Sun et al., 2023a), SigLIP2 (Tschannen et al., 2025), and Sample4Geo (Deuser et al., 2023). For oracle grounding, we compare with OCGNet (Huang et al., 2025), DetGeo (Sun et al., 2023b), Sample4Geo (Deuser et al., 2023), TROGeo (Zhang & Zhu, 2025), and SMGeo (Zhang et al., 2025). UnifyGeo (Song et al., 2025) and TransGeo (Zhu et al., 2022) are adapted to return both retrieval rankings and boxes. Adaptation details are provided in Appendix F.3. Suffixes R/G denote retrieval-only and grounding-only training.

Table 5: **Oracle grounding on WideUAV.** The paired tile is supplied for 89,728 queries. For the center prior, we use the nominal target extent for the known altitude.

Method	Heading	Heatmap	mIoU (%)↑	Acc@50 (%)↑	CDE (m)↓
Center prior (altitude-conditioned)	×	×	3.20	1.03	299.94
OCGNet (adapt.) (Huang et al., 2025)	×	×	6.36	3.39	340.60
DetGeo (adapt.) (Sun et al., 2023b)	×	×	7.35	5.01	323.71
Sample4Geo (adapt.) (Deuser et al., 2023)	×	×	10.39	8.73	322.13
TROGeo (adapt.) (Zhang & Zhu, 2025)	×	×	11.24	9.49	294.69
SMGeo (adapt.) (Zhang et al., 2025)	×	×	11.28	6.57	288.50
Grounding-only variants					
UniSAG-G			21.28	23.82	289.51
UniSAG-G + heatmap	×	✓	16.53	17.14	290.79
UniSAG-G + heading	✓	×	27.71	32.19	245.70
UniSAG-G + heading + heatmap	✓	✓	**34.40**	**44.07**	**212.72**
UniSAG, full	✓	✓	30.84	36.77	238.05

Table 6: **Retrieval-gated geo-grounding.** uAcc@25 counts all queries after gating; uCDE is conditional on retrieval success. A dash denotes an unavailable metric.

Method	R@1 (%)↑	uIoU (%)↑	uAcc@25 (%)↑	uCDE (m)↓
UnifyGeo (adapt.) (Song et al., 2025)	28.84	1.73	2.99	335.27
TransGeo (adapt.) (Zhu et al., 2022)	45.67	3.80	6.43	324.75
Grounding → retrieval	56.84	5.05	9.10	392.80
Retrieval → grounding	38.36	10.53	15.42	247.77
Without heading	57.15	11.10	17.28	274.32
Separate-encoder variant	54.99	17.48	25.52	229.11
UniSAG, full	**65.75**	**22.67**	**32.73**	**213.61**

5.1 Main Results

Tile retrieval. UniSAG-R obtains 62.03% R@1, exceeding OpenCLIP, the strongest compared baseline, by 17.55 percentage points (Table 4). Joint training raises R@1 to 65.75%, a gain of 3.72 points, and also improves R@5 and R@10.

Oracle grounding. Heading and heatmap supervision jointly improve oracle grounding (Table 5). Heading conditioning raises UniSAG-G from 21.28% to 27.71% mIoU, and adding heatmap supervision reaches 34.40%. Heatmap supervision alone lowers mIoU to 16.53%, indicating that its benefit depends on the accompanying geometric conditioning in this comparison. The jointly trained model achieves 30.84% mIoU, trading some within-tile accuracy for improved retrieval. OCGNet and DetGeo could be particularly affected by off-center targets and the loss of target detail when wide tiles are resized.

Retrieval-gated cross-view geo-grounding. The complete model obtains $22.67 \pm 0.43\%$ uIoU, $32.73 \pm 1.01\%$ uAcc@25, and 213.61 ± 4.95 m uCDE (Table 6). Adapted UnifyGeo and TransGeo reach 1.73% and 3.80% uIoU, respectively. Removing heading reduces uIoU to 11.10%. Setting oracle IoU to zero on retrieval failures lowers the aggregate score from 30.84% oracle mIoU to 22.67% uIoU.

5.2 Ablation Studies

Training configurations. We compare joint training with separate-encoder and sequential configurations in Table 6. The separate-encoder variant also has a different objective balance and text-input configuration; it obtains 54.99% R@1 and 17.48% uIoU. Both sequential schedules underperform joint training on the coupled task, reaching 5.05% uIoU for grounding followed by retrieval and 10.53% for the reverse order.

Loss-weight sensitivity. The full model obtains the highest mean R@1 and uIoU among the configurations in Table 7. At $\lambda_{\text{box}} = 0.9$, oracle mIoU is higher and CDE is lower, while R@1



Figure 4: **Selected qualitative results on WideUAV.** Each row contains a UAV query, top-1 retrieval, oracle grounding on the paired tile, and top-2, top-3, and top-4 retrievals. Green/red top-1 borders denote success/failure. Scores are maximum regional cosine similarities to the query. Oracle grounding is correct on the paired tile even when retrieval fails.

Table 7: **Ablation on λ_{box} over 3 runs.** We fix $\lambda_{\text{hm}} = 0.01$.

λ_{box}	R@1 (%)↑	mIoU (%)↑	uIoU (%)↑	CDE (m)↓	uCDE (m)↓
0.1	61.63 ±0.73	27.16 ±1.19	19.46 ±1.04	266.50 ±6.05	233.58 ±5.24
0.3	62.91 ±0.43	28.72 ±1.28	20.60 ±0.94	247.49 ±6.59	218.78 ±5.03
0.5	**65.75** ±0.42	30.84 ±0.48	**22.67** ±0.43	238.05 ±6.12	**213.61** ±4.95
0.7	61.38 ±0.42	30.18 ±0.55	20.94 ±0.52	243.05 ±4.19	217.93 ±4.21
0.9	57.27 ±0.69	**32.29** ±0.65	20.56 ±0.28	**236.24** ±5.03	214.52 ±5.97

Table 8: **Ablation on λ_{hm} over 3 runs.** We fix $\lambda_{\text{box}} = 0.5$.

λ_{hm}	R@1 (%)↑	mIoU (%)↑	uIoU (%)↑	CDE (m)↓	uCDE (m)↓
0.01	**65.75** ±0.42	**30.84** ±0.48	**22.67** ±0.43	**238.05** ±6.12	**213.61** ±4.95
0.05	64.45 ±0.53	30.33 ±0.70	22.03 ±0.49	239.94 ±5.61	214.04 ±5.02
0.1	63.71 ±0.51	29.32 ±0.17	21.11 ±0.17	249.86 ±2.92	223.00 ±2.46
0.2	62.44 ±0.75	28.59 ±0.51	20.25 ±0.36	252.35 ±3.67	224.65 ±3.86

falls to 57.27%, illustrating the trade-off between retrieval and within-tile grounding across these configurations.

Heatmap-loss weight. Reducing λ_{hm} from 0.2 to 0.01 raises mean R@1 from 62.44% to 65.75% and improves all grounding metrics (Table 8). Within the tested range, $\lambda_{\text{hm}} = 0.01$ yields the best mean for every reported metric.

Performance across acquisition conditions. Performance improves with altitude across the acquisition grid (Table 9): mean oracle mIoU rises from 19.00% at 150 m to 41.87% at 300 m, and R@1 rises from 54.87% to 73.33%. Variation across headings within each altitude is smaller. Altitude, visible context, and bbox size change together, so this comparison does not isolate their individual effects.

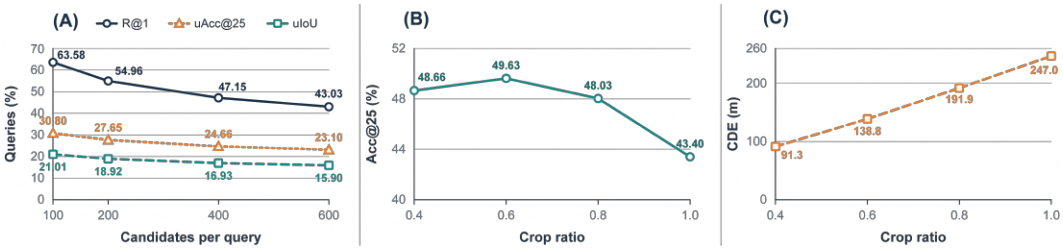

Figure 5: **Scaling retrieval candidates and tile extent.** We evaluate one fixed checkpoint in all panels. (A) Retrieval-gated performance as the candidate set grows; each set contains the paired tile and sampled negatives. (B–C) Oracle Acc@25 and CDE as the crop expands to the full test-tile footprint of $1183\,\mathrm{m} \times 660\,\mathrm{m}$ at a fixed input resolution. Distance conversion follows each crop footprint. For the reduced crops in (B–C), crop positions are chosen to retain the annotated target; protocol details are provided in Appendix D.2.2.

Table 9: **Retrieval and grounding across acquisition conditions over 3 runs.** Each altitude–heading condition contains 2,804 queries.

Altitude	0°	45°	90°	135°	180°	225°	270°	315°
Retrieval R@1 (%)								
150 m	55.22 ± 0.66	54.76 ± 1.02	54.80 ± 1.21	54.43 ± 0.66	55.47 ± 0.38	54.41 ± 0.89	54.67 ± 1.76	55.23 ± 0.91
200 m	65.06 ± 0.44	64.71 ± 0.45	63.92 ± 0.57	64.35 ± 0.88	65.47 ± 1.49	64.04 ± 0.46	64.41 ± 0.73	64.44 ± 0.47
250 m	71.21 ± 0.05	69.95 ± 0.50	70.92 ± 1.10	69.69 ± 0.77	70.05 ± 0.46	69.97 ± 0.68	70.33 ± 0.83	69.92 ± 0.79
300 m	**73.69** ± 0.25	**73.94** ± 0.77	**73.42** ± 0.75	**73.89** ± 0.45	**72.29** ± 0.55	**73.41** ± 0.73	**72.54** ± 0.29	**73.49** ± 0.95
Oracle mIoU (%)								
150 m	19.32 ± 0.55	19.25 ± 0.77	19.29 ± 0.46	19.00 ± 0.51	19.20 ± 0.49	18.41 ± 0.62	19.00 ± 0.89	18.58 ± 0.48
200 m	27.88 ± 1.07	27.29 ± 1.15	27.81 ± 0.71	26.63 ± 1.00	27.66 ± 0.45	26.07 ± 0.83	27.42 ± 0.66	26.49 ± 0.64
250 m	36.25 ± 0.74	35.33 ± 0.50	36.53 ± 0.44	34.88 ± 0.99	35.97 ± 0.70	33.94 ± 1.03	35.28 ± 1.30	34.40 ± 0.88
300 m	**42.46** ± 0.55	**41.93** ± 0.37	**42.97** ± 0.43	**41.21** ± 0.78	**42.08** ± 0.62	**40.88** ± 1.06	**42.11** ± 0.81	**41.30** ± 0.75

5.3 Retrieval Difficulty and Error Analysis

Candidate-set scaling. With one fixed checkpoint, increasing the candidate set from 100 to 600 tiles lowers R@1 from 63.58% to 43.03%, uIoU from 21.01% to 15.90%, and uAcc@25 from 30.80% to 23.10% (Figure 5A). The grounding inputs remain fixed, so this decline reflects the greater retrieval difficulty introduced by additional distractors.

Satellite-extent scaling. At a fixed input resolution, broader crops provide more context at the cost of ground detail (Figure 5B–C). As extent expands from approximately $473\,\mathrm{m} \times 264\,\mathrm{m}$ to the full test-tile footprint of $1183\,\mathrm{m} \times 660\,\mathrm{m}$, oracle CDE for the same fixed checkpoint rises from 91.3 to 247.0 m. Acc@25 peaks at 49.63% for a crop ratio of 0.6 before falling to 43.40% at full test-tile extent. Reduced test crops are confined to this extent experiment; the main evaluations retain full test tiles (Appendix D.2.2).

Center prior and retrieval-conditioned errors. Placing the nominal altitude-dependent box at the tile center gives only 3.20% mIoU (Table 5). Retrieval also discards some queries with accurate paired-tile grounding; we analyze these coupled errors and shortlist headroom in Appendix D.1, with qualitative examples in Appendix D.3.

6 Conclusion

We introduce *cross-view geo-grounding*, which extends UAV-to-satellite localization from tile retrieval to joint retrieval and within-tile target grounding. WideUAV provides a benchmark for this setting with wide-area satellite imagery, target annotations, and controlled UAV viewpoints, while UniSAG unifies retrieval and geometry-conditioned grounding through shared visual representations. Our experiments show that the two components interact asymmetrically: grounding supervision strengthens retrieval representations, whereas retrieval supervision contributes little to grounding. Overall, the results indicate that identifying the relevant satellite region is only part of wide-area cross-view localization and accurately resolving the target within that region remains a distinct challenge.

DISCLOSURE OF AI USE

We use generative AI tools to assist with literature searches, code implementation and debugging, figure preparation, and manuscript drafting. We also use model-generated target descriptions in the auxiliary text-supervision study. The authors are responsible for reviewing and validating all AI-assisted contributions and take full responsibility for the final manuscript, including its methods, results, and conclusions.

REFERENCES

Ming Dai, Enhui Zheng, Zhenhua Feng, Lei Qi, Jiedong Zhuang, and Wankou Yang. Vision-based uav self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*, 33:493–508, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16801–16810. IEEE, 2023.

Oussema Dhaouadi, Riccardo Marin, Johannes Meier, Jacques Kaiser, and Daniel Cremers. Ortholoc: Uav 6-dof localization and calibration using orthographic geodata. *Advances in Neural Information Processing Systems*, 38, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7258–7267. IEEE, 2018.

Zheyang Huang, Jagannath Aryal, Saeid Nahavandi, Xuequan Lu, Chee Peng Lim, Lei Wei, and Hailing Zhou. Object-level cross-view geo-localization with location enhancement and multi-head cross attention. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

Yannis Kalantidis, Mert Bulent SARIYILDIZ, Rafael Rezende, Philippe Weinzaepfel, Diane Larlus, and Gabriela Csurka. Weatherproofing retrieval for localization with generative ai and geometric consistency. In *International Conference on Learning Representations*, volume 2024, pp. 9199–9235, 2024.

Vincent Leroy, Jerome Revaud, Thomas Lucas, and Philippe Weinzaepfel. Win-win: Training high-resolution vision transformers from two windows. In *International Conference on Learning Representations*, volume 2024, pp. 48749–48767, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Valentinos Pariza, Mohammadreza Salehi, Gertjan J Burghouts, Francesco Locatello, and Yuki Asano. Near, far: Patch-ordering enhances vision foundation models' scene understanding. In *International Conference on Learning Representations*, volume 2025, pp. 101552–101579, 2025.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Congpei Qiu, Yanhao Wu, Wei Ke, Xiuxiu Bai, and Tong Zhang. Refining clip's spatial awareness: A visual-centric perspective. In *International Conference on Learning Representations*, volume 2025, pp. 14450–14476, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Xuelun Shen, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, Cheng Wang, et al. Gim: Learning generalizable image matcher from internet videos. In *International Conference on Learning Representations*, volume 2024, pp. 13182–13197, 2024.

Zhuo Song, Ye Zhang, Kunhong Li, Longguang Wang, and Yulan Guo. A unified hierarchical framework for fine-grained cross-view geo-localization over large-scale scenarios. *arXiv preprint arXiv:2505.07622*, 2025.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023a.

Yuxi Sun, Yunming Ye, Jian Kang, Ruben Fernandez-Beltran, Shanshan Feng, Xutao Li, Chuyao Luo, Puzhao Zhang, and Antonio Plaza. Cross-view object geo-localization in a local region with satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023b.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *International conference on learning representations*, volume 2024, pp. 52484–52503, 2024.

Wenjia Xu, Yaxuan Yao, Jiaqi Cao, Zhiwei Wei, Chunbo Liu, Jiuniu Wang, and Mugen Peng. Uav-visloc: A large-scale dataset for uav visual localization. *arXiv preprint arXiv:2405.11936*, 2024.

Yibin Ye, Xichao Teng, Shuo Chen, Leqi Liu, Kun Wang, Xiaokai Song, and Zhang Li. Exploring the best way for uav visual localization under low-altitude multi-view observation condition: a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1731–1741, 2026.

Yang You, Yixin Li, Congyue Deng, Yue Wang, and Leonidas Guibas. Multiview equivariance improves 3d correspondence understanding with minimal feature finetuning. In *International Conference on Learning Representations*, volume 2025, pp. 58081–58094, 2025.

Fan Zhang, Haoyuan Ren, Fei Ma, Qiang Yin, and Yongsheng Zhou. Smgeo: Cross-view object geo-localization with grid-level mixture-of-experts. *arXiv preprint arXiv:2511.14093*, 2025.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Qingwang Zhang and Yingying Zhu. Breaking rectangular shackles: Cross-view object segmentation for fine-grained object geo-localization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8197–8206. IEEE, 2025.

Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pp. 1395–1403, 2020.

Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, and Wenbo Hu. Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4825–4839, 2023.

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. IEEE, 2021.

Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1152–1161. IEEE, 2022.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

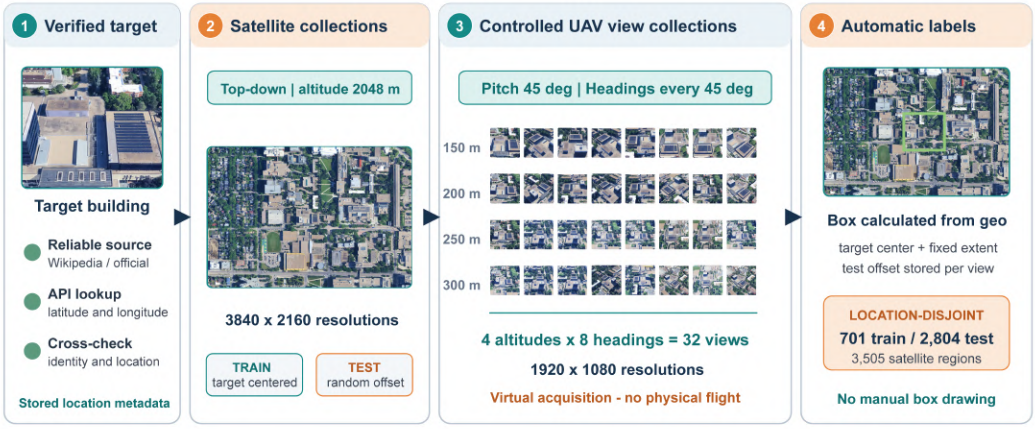

Figure 6: **Automated construction of WideUAV.** Target selection and coordinate verification are followed by Google Earth virtual-camera acquisition of wide satellite tiles and a controlled 4-altitude $\times$ 8-heading query grid. We calculate the bboxes from the UAV altitude and the target-center offset relative to the satellite-tile center. We apply random offsets to test tiles, and all views of a target remain in one split. Disjointness refers to target IDs, not necessarily to non-overlapping geographic footprints. The same workflow can be reused to add new locations.

A WideUAV Construction Details

A.1 Geographic Locations and Scene Collection

WideUAV builds upon the geographic locations of University-1652 (Zheng et al., 2020). We re-collect imagery for its 72 university scenes and extend the geographic scope with 134 additional universities and 13 city scenes. The resulting benchmark covers 206 universities and 13 city scenes across six continents, 33 countries, and 222 city labels. For each target location, we define a designated geographic coordinate and render both satellite and UAV imagery using the same Google Earth rendering pipeline. Using a shared source reduces appearance inconsistencies caused by combining imagery from different mapping providers while allowing the satellite and UAV cameras to be configured independently. The construction procedure is in Figure 6.

A.2 Satellite Image Acquisition

For each target location, we first render a wide top-down satellite image. During acquisition, the virtual satellite camera is placed at an altitude of 2,048 m, producing a 3840×2160 raw image with a maximum footprint of approximately $2370\,\mathrm{m} \times 1328\,\mathrm{m}$. For evaluation, the satellite region covers approximately $1183\,\mathrm{m} \times 660\,\mathrm{m}$. The target is not constrained to the image center; instead, the tile center is spatially offset from the designated target location. These offsets produce targets distributed across the interior of the satellite frame and introduce within-tile spatial uncertainty. For model input, the full evaluation tile is resized by a factor of five along each spatial dimension, yielding a resolution of 768×432. This corresponds to approximately 1.534 m per pixel. We therefore distinguish the native acquisition resolution from the model input resolution throughout the paper.

A.3 UAV Query Acquisition

Each target location is associated with 32 UAV queries generated from a fixed grid of four altitudes and eight headings. The UAV altitudes are 150, 200, 250, and 300 m, while headings are sampled every $45°$ over $360°$. All queries use a fixed pitch of $45°$. This acquisition protocol disentangles altitude and heading variation: one factor changes while the other is controlled. In contrast, the UAV imagery in University-1652 is collected along a spiral trajectory in which altitude and viewing direction vary jointly (Zheng et al., 2020). The fixed acquisition grid in WideUAV allows the effects of altitude and heading to be analyzed independently, as reported in Table 9. The oblique UAV views preserve scene context in front of the camera and are intended to reflect forward-looking observations rather than assuming a continuously nadir-facing platform.

A.4 Target Bounding-Box Annotation

Each satellite tile contains one designated target corresponding to the geographic location observed at the center of the UAV query. We derive its axis-aligned bounding box geometrically rather than manually tracing a building footprint. Let the designated target coordinate be offset from the satellite-tile center by a known geographic displacement. We first project this displacement into image coordinates to determine the bounding-box center. The box extent is then determined by the UAV altitude, producing an altitude-conditioned target region around the designated coordinate.

Before clipping at image boundaries, the nominal bounding-box side lengths on the native satellite canvas are 330, 414, 528, and 678 pixels for UAV altitudes of 150, 200, 250, and 300 m, respectively. These correspond to approximate physical extents of 101, 127, 162, and 208 m and occupy 1.31%, 2.07%, 3.36%, and 5.54% of the native tile area.

The nominal target extent is therefore fixed for each altitude. If a target lies sufficiently close to the satellite boundary, the corresponding bounding box is clipped to the valid image region. During training, altitude-conditioned boxes undergo the same spatial transformations as their associated satellite crops. The resulting annotations should be interpreted as altitude-dependent target neighborhoods around designated geographic coordinates rather than manually traced semantic building boundaries.

A.5 Target Position Distribution

To prevent the satellite retrieval problem from collapsing into center matching, target locations are spatially displaced from the satellite-tile center. Over the valid test annotations, the mean displacement from the tile center is 299.94 m and the median displacement is 291.78 m. Figure 2D visualizes the resulting target-center distribution in the satellite coordinate frame. The spread across the tile interior ensures that successful localization requires prediction of the target position rather than assuming a fixed image-center prior.

A.6 Train–Test Partition

All 32 UAV observations associated with the same target location are assigned to the same partition. The training set contains 701 target locations and 22,432 UAV queries, while the test set contains 2,804 target locations and 89,728 UAV queries. The split is target-ID-disjoint. It is not designed to be strictly campus-disjoint or geographically non-overlapping: geographically nearby locations may therefore appear in different partitions. We retain this property because the benchmark focuses on target-level cross-view retrieval and grounding rather than geographic-domain generalization. The geographic proximity between training and test locations is quantified in Appendix B.3.

B Evaluation Protocol and Benchmark Validity

We first establish the scoring rules, statistical uncertainty, and geographic scope of the benchmark. We then examine changes in query appearance and candidate ambiguity (Appendix C), separate retrieval and grounding errors (Appendix D), and assess two distinct roles for language in this task (Appendix E). Implementation details are collected in Appendix F.

Scope of the evidence. The protocol, geographic, joint-failure, and shortlist analyses concern the three full-model runs on all 89,728 queries, retaining their 100-candidate retrieval decisions and paired-tile oracle predictions. The synthetic-appearance experiments evaluate the same three checkpoints on transformed queries. Fixed-city and coverage diagnostics instead draw on earlier fixed-checkpoint evaluations. Auxiliary text supervision is an earlier one-run-per-variant comparison, and the VLM pilot contains one run per configuration on 32 locations. We keep these populations and training configurations distinct when interpreting their results.

B.1 Accepted tiles, exact pairs, and validation-list overlap

Each candidate list contains the paired tile and 99 other tiles sampled without replacement. Additional positive tiles listed in the inclusion metadata are accepted when sampled. We keep the ordered query lists and candidate lists fixed across protocol comparisons.

Table 10: **Sensitivity to the acceptance rule and query exclusions.** Values are averaged over the same three full-model runs. The exclusion row changes the query population only; candidate lists and trained weights are unchanged. All performance values are percentages.

Protocol	Queries	R@1	Oracle mIoU	Gated uIoU
Inclusion-aware, main protocol	89,728	65.75	30.84	22.67
Exact paired tile	89,728	61.38	30.84	20.87
Inclusion-aware, excluding validation IDs	89,152	65.77	30.80	22.62

Table 11: **Target-cluster uncertainty for the current full model.** The estimate is averaged over three runs; 95% percentile intervals resample the 2,804 target IDs. Values are percentages.

Metric	Estimate	95% interval
R@1	65.75	[64.71, 66.83]
R@5	90.57	[89.83, 91.32]
Oracle mIoU	30.84	[29.92, 31.76]
Oracle Acc@50	36.77	[35.53, 37.99]
Gated uIoU	22.67	[21.85, 23.46]
Gated uAcc@25	32.73	[31.60, 33.84]
Gated uAcc@50	27.25	[26.18, 28.29]

With the candidate lists and paired-tile oracle boxes fixed, we compare inclusion-aware acceptance with exact-pair scoring (Table 10). Requiring the top-1 ID to equal the paired-tile ID lowers R@1 by 4.38 percentage points and gated uIoU by 1.80 points.

The archived split audit contains 20 validation IDs in the raw test list; 18 occur in the effective 2,804-target evaluation. Removing their 576 queries changes R@1 from 65.75% to 65.77% and uIoU from 22.67% to 22.62%. These post hoc exclusions leave any earlier model selection involving those locations unchanged. Training and effective test target IDs remain disjoint, and the fixed-city queries have no training or validation IDs.

Distance units. CDE is converted from pixels to meters using the tile footprint. For a $1183\,\mathrm{m} \times 660\,\mathrm{m}$ tile resized to 768×432, the horizontal and vertical scales are 1.540 and 1.528 m/pixel. We use their geometric mean, giving $\mathrm{CDE_m} = 1.534\ \mathrm{CDE_{px}}$; we apply the corresponding scales for coverage crops.

B.2 Uncertainty with repeated views

Each target contributes 32 correlated views. We resample target IDs while retaining all views of each selected target, first averaging target-level statistics over the three runs and then computing 10,000 bootstrap replicates. The percentile intervals in Table 11 measure target-sampling uncertainty for these trained models, treating targets as independent clusters. Across-run standard deviations are reported separately in the main paper.

B.3 Spatial Proximity Between Training and Test Tiles

Target-ID disjointness does not imply geographic separation. We assess geographic proximity by matching the 2,804 evaluated test IDs to the archived KML inventory and computing great-circle distance to the nearest of the 701 training-tile centers. Distances are measured between shifted test-tile camera centers and raw training-tile camera centers. Thirty test-tile centers (1.07%) lie within 100 m of a training-tile center, 142 (5.06%) within 500 m, and 233 (8.31%) within 1 km. The minimum center distance is 18.54 m. The 2,478 targets at least 5 km from every training-tile center retain 64.85% R@1 and 21.11% uIoU (Table 12). Several nearby groups achieve higher scores, although the groups also differ in city, acquisition source, and scene appearance.

C Robustness to Query Appearance and Candidate Ambiguity

With the evaluation scope established, we examine two sources of difficulty: changes in the UAV observation and visual ambiguity among satellite candidates. The appearance experiments retain

Table 12: **Performance stratified by proximity to training tiles.** Distances are between KML tile-camera centers. Metrics are averaged over three full-model runs under the unchanged sampled-candidate protocol. Each target contributes 32 queries. Performance values are percentages.

Nearest training-tile center	Targets	R@1	Oracle mIoU	Gated uIoU
$<$ 250 m	70	82.40	42.62	34.97
250–500 m	72	69.98	42.45	30.69
500 m–1 km	91	69.56	49.94	36.92
1–5 km	93	70.18	46.35	34.61
$\geq$ 5 km	2,478	64.85	28.88	21.11



Figure 7: **Data card for synthetic UAV appearance variants.** Each row contains the same query view under all five transformations, with the original in the first column. Both examples have altitude 150 m and heading 0°. The variants change color, exposure, and visibility while retaining the spatial layout.

the main candidate lists and checkpoints; the city-gallery diagnostic instead evaluates neighboring tiles under exact-pair scoring. The two settings probe different conditions and are not a controlled comparison with each other.

C.1 Weather, Seasonal Appearance, and Illumination

We next examine how changes in UAV appearance affect retrieval and grounding while keeping satellite tiles and target extents fixed.

Setup. We create five synthetic variants of each of the 89,728 test queries, yielding 448,640 additional images, and evaluate the three full-model checkpoints without adaptation. Within each run, query order, 100-candidate lists, altitude, heading, and target-extent annotations remain identical across conditions. Input resolution and scoring follow the main protocol.

Rain and snow add particle overlays with cooler color tones; fog reduces contrast through a smoothly varying transmission field. Autumn shifts the hue of vegetation candidates identified from color and texture, while low light reduces exposure in linear color space and adds weak noise. Each transformed image is blended with its original at strength 0.65, preserving the image coordinates.

Results. Under the transformations illustrated in Figure 7, rain, snow, and fog reduce R@1 by 3.19–5.18 percentage points, with oracle mIoU remaining between 29.39% and 30.16% (Table 13). Autumn produces a larger retrieval decline to 51.15% R@1, while oracle mIoU remains at 28.15%. Low light affects both stages most strongly, reducing R@1 to 50.64%, oracle mIoU to 24.81%, and gated uIoU to 14.18%.

C.2 Fixed City Galleries and Selective Prediction

Complementing the appearance perturbations above, we assess candidate ambiguity with 13 fixed galleries of neighboring satellite tiles. This diagnostic concerns one earlier fixed checkpoint rather than the three full-model runs. All 32 views of each of the 1,164 query targets are evaluated. The

Table 13: **Sensitivity to UAV appearance.** Values are mean ± sample standard deviation over three runs on the complete test split, in percent. The original-image row corresponds to the full model.

Query appearance	R@1	R@5	Oracle mIoU	Gated uIoU
Original	$\mathbf{65.75} \pm 0.42$	$\mathbf{90.57} \pm 0.21$	$\mathbf{30.84} \pm 0.48$	$\mathbf{22.67} \pm 0.43$
Rain	62.34 ± 0.32	89.56 ± 0.20	29.39 ± 0.09	20.42 ± 0.15
Snow	60.57 ± 0.36	88.29 ± 0.18	29.53 ± 0.15	19.87 ± 0.27
Fog	62.56 ± 0.65	89.34 ± 0.38	30.16 ± 0.72	21.07 ± 0.66
Autumn	51.15 ± 1.34	80.40 ± 1.17	28.15 ± 0.59	16.60 ± 0.59
Low light	50.64 ± 2.43	82.60 ± 1.63	24.81 ± 0.95	14.18 ± 1.06

Table 14: **Results in fixed neighboring city galleries.** Gallery size is the number of tiles; performance values are percentages. Micro averages weight queries equally, and macro averages weight the 13 cities equally.

City	Gallery	Queries	R@1	uIoU	uAcc@50
Atlanta	99	3,168	24.24	6.82	7.20
Belo Horizonte	110	3,520	9.83	1.40	1.59
Budapest	77	2,464	13.92	2.00	2.11
Dallas	99	3,168	20.77	6.37	7.13
Denver	88	2,816	25.32	7.27	8.56
Indianapolis	88	2,816	32.32	7.80	7.95
Johannesburg	110	3,520	26.76	4.93	5.20
Madrid	87	2,752	28.42	7.21	9.05
Milan	77	2,464	19.56	6.55	7.63
Munich	77	2,464	34.90	9.96	12.05
Prague	77	2,464	33.89	10.36	11.53
Santiago	99	3,168	19.22	3.81	4.20
Vienna	77	2,464	29.95	8.84	11.00
All queries (micro)	1,165	37,248	24.12	6.21	7.07
City average (macro)	–	–	24.55	6.41	7.32

1,165 gallery tiles include one Madrid tile whose query images are missing; it remains a distractor. Retrieval requires the exact paired tile, and grounding is performed on the actual top-1 tile.

City-level R@1 ranges from 9.83% to 34.90%, indicating substantial heterogeneity despite the modest gallery sizes. Within these 13 cities, resampling target IDs while preserving city sample counts gives a 95% interval of [22.39, 25.86]% for micro R@1 and [5.49, 6.95]% for micro uIoU.

Accuracy at reduced coverage. The city-level results quantify difficulty across locations; we next ask whether retrieval confidence identifies more reliable queries within this population. For selective prediction, we rank queries by the difference between the largest and second-largest tile similarity scores and retain those with the largest margins (Table 15). The retention thresholds are empirical quantiles of the test scores.

Larger margins identify a more accurate subset, but 37.07% of the retained queries still retrieve an incorrect tile at approximately 10% coverage.

D Retrieval and Grounding Failure Analysis

Overall scores combine errors at two stages. We first separate these outcomes on the full test split and estimate the headroom from a better retrieval shortlist. We then examine box geometry in the archived city evaluations and connect the quantitative diagnostics to selected qualitative cases.

D.1 Joint Failure Modes and Shortlist Headroom

We classify each query by accepted top-1 retrieval and paired-tile oracle IoU above 0.5 (Table 16). Correct retrieval with insufficient oracle overlap is the largest failure category, comprising 38.50% of queries. Conversely, 9.52% have an accurate oracle box but fail retrieval.

Table 15: **Selective prediction by retrieval-score margin.** The fixed-city checkpoint and predictions are unchanged. All scores are computed over the retained queries; uIoU and uAcc still assign zero to retrieval failures within that subset.

Retained fraction	Queries	R@1 (%)	uIoU (%)	uAcc@50 (%)
100%	37,248	24.12	6.21	7.07
50%	18,624	33.27	9.19	10.56
25%	9,312	45.33	13.22	15.22
10%	3,725	62.93	18.42	21.23

Table 16: **Joint retrieval and localization outcomes.** Entries are percentages of all queries, averaged over the three full-model runs, with target-cluster 95% intervals. Localization success refers to paired-tile oracle IoU > 0.5.

Top-1 retrieval	Oracle localization	Query share	95% interval
Accepted	Successful	27.25	[26.18, 28.29]
Accepted	Unsuccessful	38.50	[37.46, 39.61]
Rejected	Successful	9.52	[8.98, 10.06]
Rejected	Unsuccessful	24.73	[23.77, 25.67]

Grounding conditioned on retrieval. For the full model, oracle mIoU is 34.47% on accepted top-1 retrievals and 23.86% on retrieval failures, averaged over the three runs. The corresponding Acc@50 values are 41.45% and 27.79%. The retrieved-positive subset is itself easier to ground.

Headroom from a retrieval shortlist. To estimate shortlist headroom under ideal selection, we accept a query whenever an accepted tile appears among the top K and retain its paired-tile oracle IoU (Table 17).

The top-5 calculation recovers 6.36 points of IoU relative to top-1 gating, while expanding from five to ten candidates adds 1.00 point. Most of the estimated gain arises within the first five candidates. This calculation assumes ideal candidate selection and retains paired-tile oracle boxes; it is not an evaluated reranking system.

D.2 Proxy Extents and Coverage-Scaling Diagnostics

The joint-failure analysis identifies insufficient box overlap as a frequent failure even after correct retrieval. To examine the geometry of this error, we turn to the archived fixed-city predictions from Appendix C.2. As described in the dataset section, bbox centers are calculated from target-center offsets relative to the satellite-tile center, and extents are calculated from UAV altitude. These altitude-dependent neighborhoods motivate separate checks of center estimation, extent estimation, and the crop geometry of the coverage experiment.

D.2.1 Separating center and size errors

We separate position and size errors in the fixed-city predictions by replacing the predicted width and height with the nominal altitude-dependent side length: 66.0, 82.8, 105.6, or 135.6 pixels in the 768×432 frame. Retrieved tiles and predicted centers remain fixed, and boxes are reconstructed without clipping to preserve their centers. Two oracle controls replace either the size or the center with its ground-truth value.

Replacing learned sizes by nominal extents increases uIoU by 0.22 points, with a paired target-cluster 95% interval of [0.16, 0.27] points. The uAcc@50 change is only 0.02 points, with an interval of $[-0.08, 0.13]$. Replacing centers by their oracle values increases uIoU by 13.01 points, with an interval of [11.92, 14.08]. For this fixed checkpoint and proxy-label scheme, center localization is the larger remaining source of box error.

D.2.2 Crop protocol for the coverage experiment

The following crop diagnostic applies only to the coverage-scaling experiment; the default evaluation retains full satellite tiles. For the reduced crops in Figure 5B–C, we construct a feasible crop region

Table 17: **Ideal-selection diagnostic for retrieval shortlists.** The same saved top-K hit indicators and paired-tile IoUs are evaluated across three full-model runs. Values are percentages. The final row supplies the paired tile for every query.

Acceptance condition	Recall	Gated oracle IoU	Gated oracle Acc@50
Accepted tile at top 1	65.75	22.67	27.25
Accepted tile in top 5	90.57	29.02	34.70
Accepted tile in top 10	95.16	30.02	35.85
Paired tile always supplied	100.00	30.84	36.77

Table 18: **Position and extent diagnostics on fixed-city predictions.** Retrieval is fixed at 24.12% exact-pair R@1 for every row. Conditional IoU averages only successful retrievals; uIoU and uAcc@50 include zero-valued retrieval failures. Values are percentages.

Box construction	Conditional IoU	uIoU	uAcc@50
Learned center and size	25.75	6.21	7.07
Learned center, nominal altitude extent	26.64	6.43	7.09
Learned center, oracle size	26.64	6.43	7.08
Oracle center, learned size	79.66	19.22	23.51

from the ground-truth box and take the midpoint of the valid horizontal and vertical crop-origin intervals. This procedure keeps the target inside the crop but changes its relative position. The coverage experiment therefore jointly changes image detail, available context, and the target-position prior.

We quantify this effect on the same 1,164 city targets by reproducing each crop and placing the nominal altitude-dependent box at its center. The 4,656 target–altitude boxes correspond to all 37,248 city queries because the eight headings share the same satellite annotation. Normalized center displacement is the Euclidean distance from the target-box center to the crop center after scaling both image axes to unit length.

On this city subset, at ratio 0.4, the center prior reaches 52.32% mIoU despite having no image input, exceeding the learned 13.90% on these queries. At full coverage, the ordering reverses: the prior reaches 2.87% and learned grounding reaches 17.89%. The corresponding crop footprints are approximately 473×264, 710×396, 946×528, and 1183×660 m.

The archived crop and city records match by target ID, altitude, heading, and image paths. They reference the same checkpoint path but contain no weight hashes. At full coverage, oracle mIoU differs by 0.00128 percentage points between the two records; we retain the city evaluation for that row.

D.3 Qualitative Examples

The aggregate diagnostics separate retrieval failures from insufficient box overlap. We complement them with selected visual examples of accurate paired-tile grounding under both successful and failed retrieval. We select ten test queries from one fixed checkpoint trained with $\lambda_{\mathrm{hm}} = 0.01$ and $\lambda_{\mathrm{box}} = 0.5$ for Figure 8. From queries with oracle IoU of at least 0.6, we retain the highest-IoU view per location and retrieval outcome, then select eight correct exact-pair retrievals with large top-1–top-2 score margins and two urban retrieval failures. Each selected query depicts a different location. We retain the original 100-candidate lists and full satellite tiles. These deliberately selected cases illustrate different outcomes rather than their frequency in the test set.

The successful cases span sports facilities, industrial structures, treatment basins, railway infrastructure, rural buildings, a waterfront, and a wooded hillside. Their top-1–top-2 score margins range from 0.276 to 0.370, with oracle IoU from 0.667 to 0.927. In the two urban failure cases, the corresponding margins are 0.014 and 0.008 despite oracle IoU of 0.947 and 0.949; the paired tiles rank fourth and third, respectively. Accurate grounding on the paired tile cannot recover the target when retrieval selects an incorrect tile.

Table 19: **Crop geometry and grounding for the coverage experiment.** All rows contain the same city queries. The center prior has the nominal extent for the known altitude, scaled to the crop; learned scores are paired-tile oracle scores from the archived coverage evaluation and the full-tile city evaluation.

Crop ratio	Mean displacement	Prior mIoU (%)	Learned mIoU (%)	Prior Acc@50 (%)	Learned Acc@50 (%)
0.4	0.1325	52.32	13.90	50.67	2.33
0.6	0.2714	9.48	13.87	3.93	6.19
0.8	0.3356	3.18	18.69	0.84	20.67
1.0	0.3046	2.87	17.89	0.84	19.80



Figure 8: **Additional qualitative results on WideUAV.** Each row contains one case, with columns ordered as UAV query, top-1 retrieval, oracle grounding on the paired tile, and top-2, top-3, and top-4 retrievals. The first seven rows contain correct top-1 retrievals; the last row contains an urban retrieval failure. Green/red top-1 borders denote success/failure. Green and orange boxes denote ground-truth and predicted extents. Scores are maximum regional cosine similarities to the query. Oracle grounding is on the paired tile even when retrieval fails.

E Language Supervision and Multimodal Retrieval

We examine two distinct ways to incorporate language-related models: auxiliary text supervision during task-specific training, and direct candidate scoring by off-the-shelf VLMs at inference. The first comparison concerns visual-only inference after training with descriptions; the second measures retrieval accuracy and compute for general multimodal models. Their training regimes, query populations, and purposes differ, so their scores should not be interpreted as a shared ablation.

Table 20: **Auxiliary text supervision in an earlier configuration.** Both variants are evaluated on the same 89,728 queries, 2,804-tile source gallery, and 100-candidate protocol, with full test tiles at 768×432 resolution. Each row is one training run. Values are percentages.

Historical training variant	R@1	R@5	Oracle mIoU	Gated uIoU
Without auxiliary text	62.02	87.47	28.86	20.47
Fixed auxiliary text weight, 0.01	61.30	86.78	26.28	18.91

E.1 Auxiliary Text Supervision

An earlier training comparison examines auxiliary image–text alignment with generated target descriptions. Both variants perform inference from visual inputs; Table 20 contains their results.

The paired target-cluster analysis gives a text-minus-no-text R@1 difference of -0.72 points, with a 95% interval of $[-1.36, -0.09]$, and an oracle mIoU difference of -2.57 points, with an interval of $[-3.35, -1.79]$. The runs also differ in training entry point and have unrecorded training random states, limiting attribution to the auxiliary loss.

The associated training-caption inventory contains 14,020 descriptions for 701 target IDs, with 13,513 distinct strings. At least one of the relative spatial terms *left, right, top, bottom, upper, lower, foreground*, and *background* occurs in 53.05% of descriptions, and repeated views of a target can share language while differing in heading. These properties motivate checking description specificity and view consistency. Prompt-search scores selected through repeated evaluation on test locations are excluded from this comparison.

E.2 Vision–Language Model Retrieval Pilot

We next consider language at inference through off-the-shelf multimodal models, separately from the auxiliary training objective above. We evaluate six quantized vision–language models (VLMs) on 32 test locations, with one UAV view per location and each of the 32 altitude–heading combinations represented once. The same frozen 100-candidate list is retained for each query across all models. UAV images are 256×256 pixels, and full satellite tiles are resized to 768×432 pixels without cropping. This pilot measures satellite-tile retrieval; it does not evaluate grounding or general multimodal reasoning.

Scoring and execution. For each UAV–satellite pair, we ask whether the satellite tile contains the geographic location at the center of the UAV image. All 100 candidates are scored by $\log P(\mathrm{Yes}) - \log P(\mathrm{No})$ at the first answer token and ranked in descending order, with ties resolved by the frozen candidate order. The same prompt is applied to all six models without WideUAV fine-tuning. We retain one complete run per model on the same 48-GB-class RTX 4090 GPU, with the same recorded llama.cpp binary and libraries, four concurrent pair requests, eight CPU threads, and an 8,192-token context per request. Prompt caching and speculative decoding are disabled. Model-dependent quantization and internal visual tokenization remain part of each measured configuration.

Accuracy and resource accounting. We report exact-pair R@1, R@5, R@10, and mean reciprocal rank (MRR) in Table 21. Online GPU occupancy is the sum of query wall-clock times while the single GPU is reserved, divided by the query count and scaled to 1,000 queries. This measure includes preprocessing and scoring requests, but excludes model loading and warm-up; it is neither an integral of GPU utilization nor a monetary price. Concurrent pair durations are not added together to estimate query latency. All six runs complete 32/32 queries without query-level failures or retries. Gemma requires two additional requests to obtain missing answer-token log probabilities; their elapsed time is included. Peak process memory is measured separately from total device memory.

The archived quantization settings are `Q4_K_M` for InternVL3.5 and MiMo, `NVFP4-MTP-Q8attn` for Qwen3.8, and `UD-Q4_K_XL` for the remaining VLMs. These labels identify the tested weight variants, not a uniform numerical format across architectures.

Observed trade-offs. Among the six measured configurations, MiMo has the lowest online GPU occupancy at 6.41 GPU-hours per 1,000 queries and retrieves 8/32 paired tiles at rank one (Figure 9). Qwen3-VL and Qwen3.8 both retrieve the paired tile first for 11/32 queries, but the online occupancy

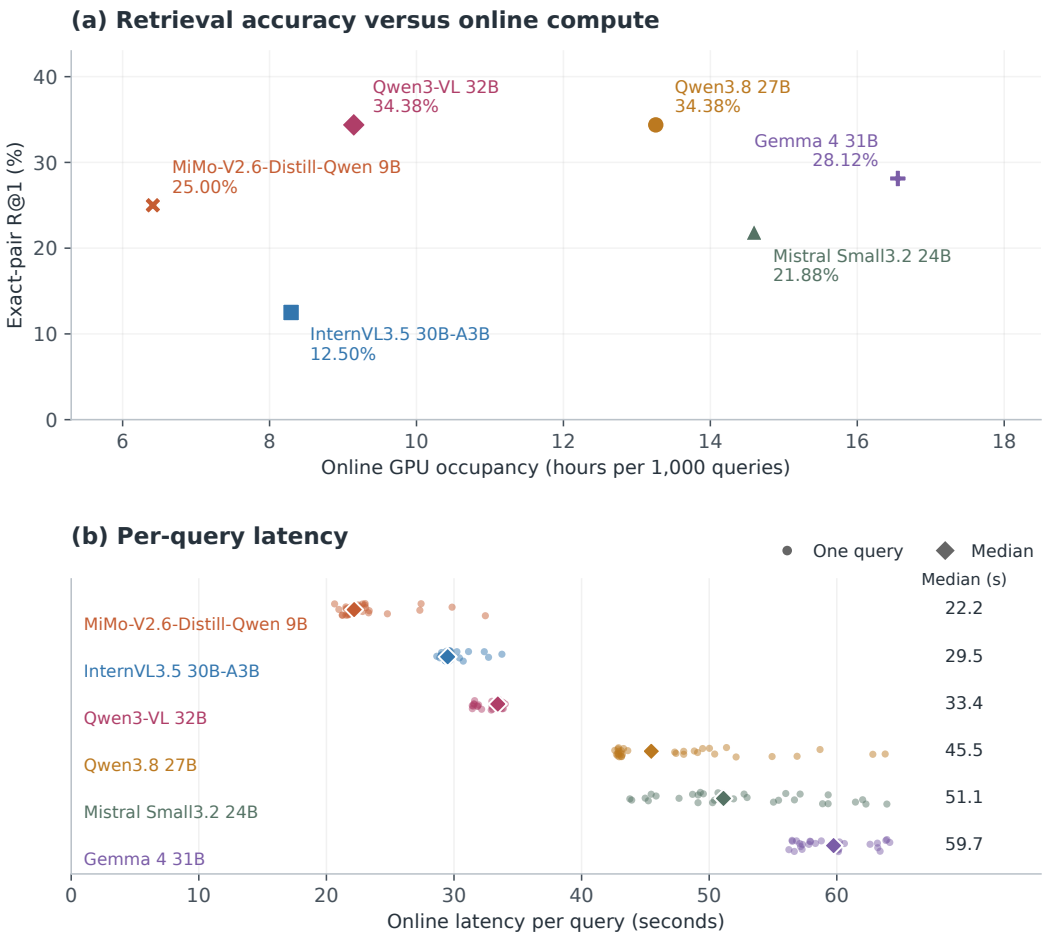

Figure 9: **Retrieval accuracy and compute on the 32-location VLM pilot.** (a) Exact-pair R@1 versus online GPU occupancy, normalized to 1,000 queries, for six VLM configurations. Labels give model names and R@1 point estimates. (b) Per-query wall-clock latency on the same 32 locations, with one dot per query and diamonds for medians. Median latency is also listed at right. Vertical jitter only separates overlapping dots.

Table 21: **Retrieval quality and resources on identical pilot queries.** Recall values are percentages; MRR is on a 0–1 scale. GPU-h denotes online occupancy per 1,000 queries, and VRAM denotes peak process memory in GiB. Bold values are best among the six VLM configurations, with lower values preferred for GPU-h and VRAM. The final row is an accuracy-only reference from one archived, WideUAV-trained UniSAG checkpoint; its original run was on a different GPU.

Model	R@1	R@5	R@10	MRR	GPU-h	VRAM
InternVL3.5 30B-A3B	12.50	46.88	62.50	0.287	8.29	21.51
Mistral Small3.2 24B	21.88	43.75	68.75	0.349	14.60	19.84
MiMo-V2.6-Distill-Qwen 9B	25.00	43.75	62.50	0.359	**6.41**	**7.79**
Gemma 4 31B	28.12	62.50	81.25	0.459	16.55	26.67
Qwen3-VL 32B	**34.38**	**78.12**	**100.00**	**0.526**	9.15	28.52
Qwen3.8 27B	**34.38**	59.38	81.25	0.482	13.26	21.13
UniSAG (FP32)	65.62	93.75	93.75	0.761	—	—

of Qwen3-VL is 31.0% lower (9.15 versus 13.26 GPU-hours per 1,000 queries). Qwen3-VL also ranks the paired tile within the first ten for all 32 queries, compared with 26/32 for Qwen3.8. The median query latency ranges from 22.2 to 59.7 seconds across the six configurations. The archived

UNISAG checkpoint retrieves 21/32 paired tiles at rank one. This reference compares a task-trained retriever with off-the-shelf VLMs, rather than controlling for training data or architecture. With only 32 locations and one run per configuration, these point estimates do not establish a statistically resolved ranking among the VLMs.

F Implementation and Reproducibility

We provide the model, loss, and baseline implementation details that supplement the main experimental setup. These settings describe the full model underlying the current three-run results; the historical diagnostics and VLM pilot retain their separately stated configurations.

F.1 Feature geometry and grounding implementation

The resulting drone and satellite feature maps have spatial sizes of 14×14 and 27×48, respectively, with 768 channels. Global attention pooling produces the drone descriptor, while a trainable attention pooler produces 15 satellite descriptors from a 3×5 partition. For geometry conditioning, the FiLM MLP receives $(\cos\theta, \sin\theta, a/300)$, with altitude normalized by the maximum acquisition height of 300 m. In the grounding branch, we detach the drone features, rotate them by bilinear spatial resampling, and average-pool them into a centered, normalized correlation kernel. A 1×1 convolution fuses the correlation logits with satellite features. To form $\hat{\mathbf{y}}'$, we resize the heatmap logits to the detector grid if needed, apply the sigmoid, broadcast over nine anchors, scale by 0.5, and detach the confidence path. Box regression is parameterized by within-cell center offsets and log size ratios relative to the selected anchor.

F.2 Loss settings

For retrieval, we set $\tau = 0.07$ and $\epsilon = 3/35$, assigning target mass 0.92 to the center-containing region and $0.08/14$ to each remaining region of the paired tile. Confidence cross-entropy has unit weight. For heatmap supervision, we set the focal modulation exponent to $\gamma = 2$. Positive and negative terms are weighted by the soft target and its complement, normalized by their respective target masses, and scaled by 15 and 45 before averaging examples. We clamp probabilities to $[10^{-4}, 1 - 10^{-4}]$ and normalization denominators to at least 10^{-6}. In Equation 4, we set $\lambda_{\text{box}} = 0.5$ and $\lambda_{\text{hm}} = 0.01$ for the full model. The heatmap-weight study keeps $\lambda_{\text{box}} = 0.5$ and evaluates λ_{hm} from 0.01 to 0.2.

F.3 Baseline implementation

The TransGeo adaptation attaches a learned convolutional box head to its query-conditioned aerial features. We implement OCGNet with a dual-encoder matcher and an anchor head, and perform Sample4Geo retrieval with regional descriptors.

F.4 Evaluation records

We retain per-query identifiers, source-file hashes, sample counts, and analysis scripts for each evaluation. Percentage conversions follow metric computation from the original fractional values. Paired bootstrap comparisons resample the same target IDs in both conditions.