GeoVLT: Camera Geometry Conditioning for UAV Vision–Language Tracking

Anonymous authors
Paper under double-blind review

Abstract

In UAV-based urban monitoring, continuous target tracking often faces changes in altitude, gimbal pitch angle and focal length, resulting in different camera geometry settings that affect target imaging and scene coverage. Most existing vision-language trackers rely on learned features to implicitly handle the resulting imaging changes, potentially reducing tracking reliability across geometry settings. Camera geometry provides physical cues about these imaging changes, enabling more reliable target modeling. Therefore, we introduce GeoVLT, a vision-language tracking framework that uses these cues throughout the tracking process. Specifically, Geometry-Conditioned Representation first adapts the target representation within a single frame. Geometry-Adaptive Temporal Propagation then maintains historical target information through multi-frame propagation. Geometry-Conditioned Decoding finally combines the single-frame representation with multi-frame historical information for coordinate prediction. To evaluate the use of camera geometry in tracking, we construct GeoUAV, a dataset that provides vision, language, and camera geometry information. For public UAV benchmarks without camera geometry, we train a lightweight estimator on GeoUAV to predict camera geometry from first-frame features. GeoVLT achieves state-of-the-art performance among the evaluated trackers, reaching 69.69% AUC on GeoUAV100 compared with 67.31% for ATCTrack, and 64.11% AUC on UAV20L-NLP compared with 61.66% for UITrack. These results show the effectiveness of explicitly using camera geometry in vision-language tracking. Our code and checkpoints are released at `https://anonymous.4open.science/r/GeoVLT`.

1 Introduction

UAVs provide aerial observations for urban monitoring and incident response tasks that require continuous tracking of ground targets. UAV vision–language tracking combines vision information with language descriptions to follow the intended target (Wang et al., 2021b; Zhou et al., 2023; Xue et al., 2025). Monitoring tasks differ in their requirements for scene coverage and target detail. These requirements lead to different flight and camera settings. Changes in flight and camera settings alter imaging conditions and affect target appearance and tracking performance. Reliable monitoring therefore requires effective tracking across these settings.

Higher altitude reduces target scale, while a longer focal length enlarges the target but narrows scene coverage. Gimbal pitch angle changes viewing direction and perspective, affecting visible target surfaces and surroundings. Figure 1(a) illustrates these imaging effects. Previous studies have shown that altitude and gimbal pitch angle affect target appearance and recognition difficulty (Wu et al., 2019; Ye et al., 2025). Reduced target detail makes similar objects harder to distinguish. Changes in scene coverage also alter the background cues available for tracking. Therefore, tracking performance depends on how these imaging differences are handled.

Existing UAV vision–language trackers rarely consider the camera geometry underlying these changes. Some methods improve target representations to handle viewpoint changes and occlusion (Li et al., 2024; Wu et al., 2025). Others use information from previous frames to improve feature extraction and refine similarity maps (Cao et al., 2022). The use of natural language descriptions further helps trackers distinguish the target from similar objects in complex UAV scenes (Xue et al., 2025). These

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.

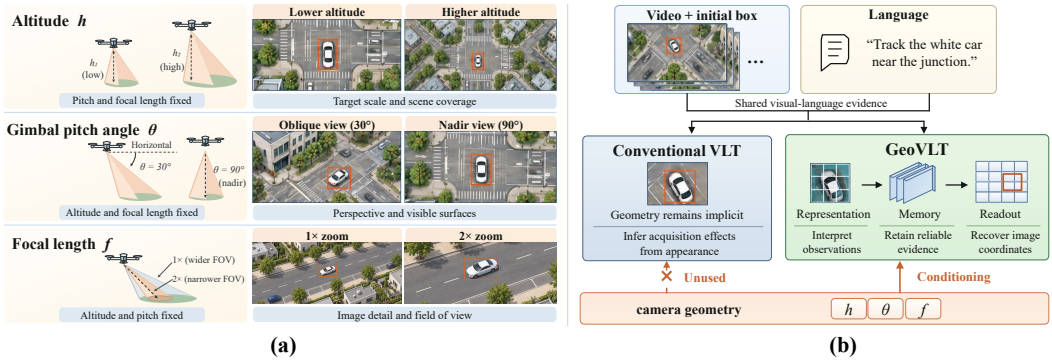


Figure 1: **Motivation.** (a) Different camera geometry settings change target imaging and scene coverage. (b) Existing vision–language trackers handle these imaging changes implicitly. GeoVLT uses camera geometry to handle them explicitly.

methods implicitly handle changes in target appearance through learned features and do not take camera geometry as an explicit input.

To address this issue, we explicitly consider camera geometry throughout tracking. First, camera geometry provides cues for interpreting target appearance in each frame through its effects on target scale, visible detail, and scene coverage. Second, camera geometry guides the retention and reuse of reliable historical target information across multiple frames. The accumulated information serves as a reference for current observations as the target moves and helps distinguish it from the background and similar objects. (Wang et al., 2021a; Xie et al., 2023; Cai et al., 2024). Finally, camera geometry provides information about the target's apparent size to guide coordinate prediction. The prediction combines current target information with historical information retained across multiple frames. (Sun et al., 2018; Zhang et al., 2020).

Accordingly, We propose GeoVLT as a vision–language tracking framework that explicitly uses camera geometry throughout tracking as illustrated in Figure 1(b). **Geometry-Conditioned Representation (GCR)** first uses camera geometry to interpret target appearance and form target features within a single frame. **Geometry-Adaptive Temporal Propagation (GATP)** then retains reliable target information across frames and reuses it under the corresponding imaging condition to support target identification as the target moves. **Geometry-Conditioned Decoding (GCD)** finally combines current target features with information retained across multiple frames. It uses camera geometry to guide coordinate prediction through cues about the target's apparent size. We also construct GeoUAV, a vision–language tracking dataset collected under multiple controlled camera geometry settings and providing the corresponding geometry information. For public UAV benchmarks without camera geometry, we design a lightweight estimator on GeoUAV to predict camera geometry from first-frame features. Our contributions are as follows:

- We introduce GeoVLT, the first framework that explicitly utilizes camera geometry information for UAV vision-language tracking. The framework incorporates geometry information throughout target representation, temporal information propagation, and coordinate prediction, enabling tracking under different imaging conditions.
- We construct GeoUAV, a UAV tracking dataset with vision, language, and camera geometry annotations under explicitly controlled geometry settings across different combinations of altitude, gimbal pitch angle, and focal length. Camera settings remain fixed within each sequence and vary across sequences.
- Experiments on GeoUAV and public UAV benchmarks show that GeoVLT achieves state-of-the-art performance among the evaluated trackers, reaching 69.69% AUC on GeoUAV100 and 64.11% AUC on UAV20L-NLP. The latter result uses camera geometry obtained from the estimator.

2 Related Work

UAV vision–language tracking. Vision–language tracking combines visual observations with language descriptions for target identification (Li et al., 2017; Wang et al., 2021b). JointNLT integrates visual grounding with tracking, while UVLTrack unifies visual and vision–language tracking through contrastive learning (Zhou et al., 2023; Ma et al., 2024). Recent methods improve language adaptation and target-context alignment through adaptive prompting, dynamic language updates, and unidirectional information modeling (Zhao et al., 2025; Li et al., 2025; Feng et al., 2025; Wang et al., 2025). UAV tracking faces substantial scale variation, viewpoint changes, and occlusion (Mueller et al., 2016; Li & Yeung, 2017; Zhang et al., 2023). TCTrack exploits temporal context, while AVTrack and ORTrack improve robustness to viewpoint changes and occlusion, respectively (Cao et al., 2022; Li et al., 2024; Wu et al., 2025). AVLTrack introduces dynamic sparse vision–language interaction for aerial tracking (Xue et al., 2025). AerialMind further studies language-guided multi-object tracking in UAV scenes (Chen et al., 2026a). These methods mainly handle appearance changes through learned visual features or language guidance. However, explicit use of camera geometry to handle changing imaging conditions remains limited.

Geometry in visual perception. Previous studies incorporate camera geometry into visual representations or estimate camera parameters from images. Metric3D introduces a canonical camera space, while CAM-Convs and ZeroDepth encode camera geometry into visual features (Yin et al., 2023; Facil et al., 2019; Guizilini et al., 2023). UniDepth predicts camera representations for feature conditioning, and Depth Pro estimates focal length from images (Piccinelli et al., 2024; Bochkovskiy et al., 2025). In general object tracking, GOT-Edit extracts 3D features from current and reference frames using a pretrained geometry model. It uses these features for online model editing (Chen et al., 2026b). These studies establish methods for representing or estimating geometric information, but do not explicitly incorporate camera geometry throughout UAV vision–language tracking. Therefore, we propose a camera-geometry-aware framework to address changing imaging conditions in UAV vision–language tracking.

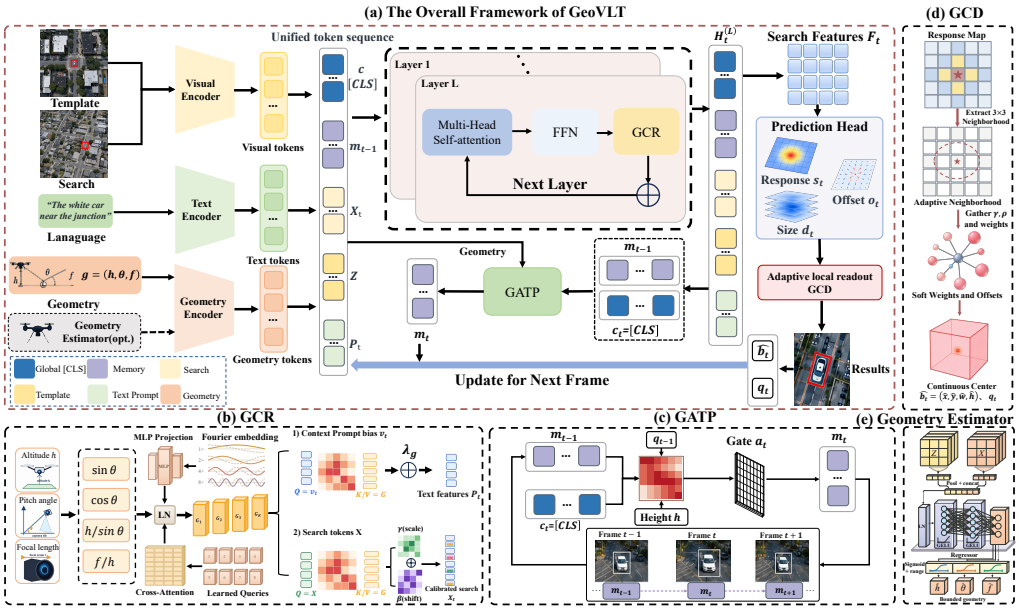


Figure 2: **Overall framework of the proposed GeoVLT.** With vision and language inputs, GCR uses camera geometry to adapt vision features and language prompts for the current frame. GATP maintains this target information across frames and uses camera geometry to control history updates. The prediction head converts features containing current-frame information and tracking history into response and offset maps. GCD then uses camera geometry to weight neighboring responses and combine their offsets for coordinate prediction. The predicted center and size form the bounding box. Distillation-Guided Geometry Estimation learns from measured camera geometry in GeoUAV. It predicts camera geometry from first-frame vision features for public datasets without this information.

3 METHODOLOGY

Overview. The GeoVLT framework shown in Figure 2 adopts a one-stream multimodal encoder (Zhao et al., 2025). The camera geometry $g = (h, \theta, f)$ consists of altitude, gimbal pitch angle, and focal length. These parameters remain fixed within each sequence and are encoded once as $G = \mathcal{E}_g(g)$. At frame t, the tracker predicts the target box $\hat{b}_t$ and its confidence score q_t while updating the recurrent memory m_t that retains target information across frames:

$$(\hat{b}_t, q_t, m_t) = \mathcal{F}_\Theta(z, x_t, \ell, g, G, m_{t-1}, q_{t-1}), \tag{1}$$

where z, x_t, and ℓ denote the template, search crop, and language description, respectively. The initial memory and confidence are set to $m_0 = 0$ and $q_0 = 1$.

Geometry encoding. We normalize h, θ, and f to their collection ranges and append $\sin\theta$, $(\cos\theta + 1)/2$, a slant-range proxy $h/\sin\theta$, and an apparent-size proxy f/h. Derived quantities are computed from camera geometry after converting angles to radians. The slant-range proxy assumes locally level ground, and both ratio-based proxies are scaled by their maxima over the collection ranges. A four-band Fourier map $\phi(u) = [u, \{\sin(2^j\pi u), \cos(2^j\pi u)\}_{j=0}^{3}]$ expands each coordinate to represent nonlinear dependencies (Tancik et al., 2020; Mildenhall et al., 2020). An MLP followed by a linear projection produces four context tokens. Eight learned queries read this context through cross-attention, and layer normalization yields $G \in \mathbb{R}^{8\times D}$. Below, $\mathcal{A}_G^{(a)}(Q)$ denotes multi-head cross-attention to G, including projections specific to branch or layer a.

3.1 GEOMETRY-CONDITIONED REPRESENTATION

GCR conditions language context and visual features on camera geometry. In the language branch, geometry augments the search context vector before prompt encoding. Let v_t denote the search context vector, C the matrix of n_p learned prompt tokens, and λ_g the coefficient controlling the geometry-conditioned context. The resulting language representation is

$$P_t = \mathcal{T}_{\text{CLIP}}\Big(\ell,\, C + \mathbf{1}_{n_p}\Big[v_t + \lambda_g \mathcal{A}_G^{(\text{lang})}(v_t)\Big]\Big), \tag{2}$$

where $\mathbf{1}_{n_p}$ broadcasts the context across prompt positions, and $\mathcal{T}_{\text{CLIP}}$ encodes the prompted description and projects the output to the visual feature width (Radford et al., 2021; Zhou et al., 2022b;a). Geometry therefore affects how the target description is contextualized before multimodal interaction.

The visual encoder receives $U_t^{(0)} = [c, m_{t-1}, X_t, Z, P_t]$, where c is a global token and X_t and Z are search and template features. At selected layers, the search feature $X_t^{(l)}$ queries the geometry tokens G. Building on feature-wise (Huang & Belongie, 2017; Perez et al., 2018) and spatially adaptive conditioning (Wang et al., 2018; Park et al., 2019; Jacenków et al., 2020), GCR applies

$$\begin{aligned}[\gamma_t^{(l)}, \beta_t^{(l)}] &= \text{Linear}_l\Big(\mathcal{A}_G^{(l)}\Big(\text{LN}(X_t^{(l)})\Big)\Big), \\ \widehat{X}_t^{(l)} &= X_t^{(l)} \odot \Big(1 + \gamma_t^{(l)}\Big) + \beta_t^{(l)},\end{aligned} \tag{3}$$

where $\gamma_t^{(l)}$ and $\beta_t^{(l)}$ are geometry-conditioned modulation parameters.

Although geometry remains fixed within a sequence, the current visual features produce different corrections through geometry conditioning. The modulated feature $\hat{X}_t^{(l)}$ continues through subsequent layers, while other tokens interact with it through self-attention.

3.2 GEOMETRY-ADAPTIVE TEMPORAL PROPAGATION

GATP propagates geometry-conditioned historical information to current search features through self-attention in each encoder layer. For one attention head, let Q_X contains search queries, K contains keys of all tokens, and V_m is the value of the memory token. Its contribution to the search attention output is

$$R_{X\leftarrow m} = \left[\text{softmax}_{\text{row}}\left(\frac{Q_X K^\top}{\sqrt{d_k}} + B_X\right)\right]_{:,m} V_m, \tag{4}$$

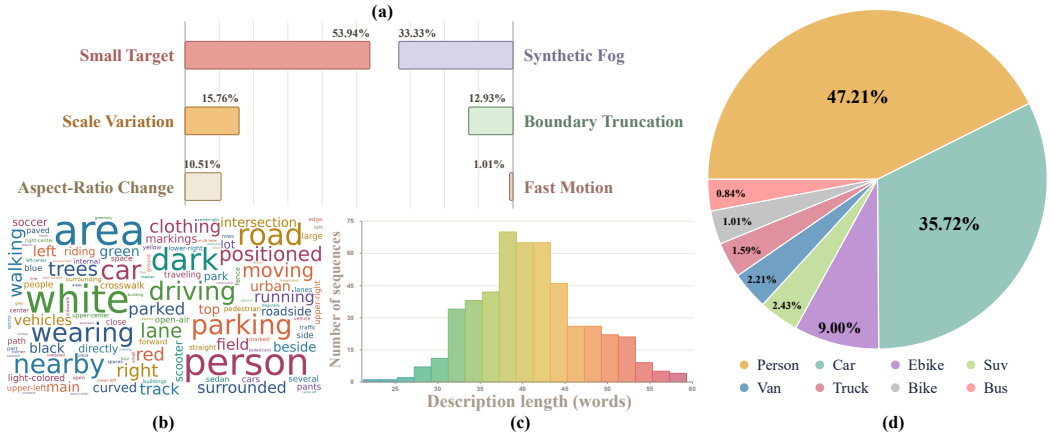


Figure 3: **GeoUAV statistics.** (a) attribute proportions, (b) description vocabulary, (c) description lengths, and (d) class distribution.

where d_k is the key dimension and B_X contains the baseline attention biases and masks. The selected column represents the attention weight assigned to memory. The normalization allows search features to balance historical information with other tokens.

After encoding, the normalized global token summarizes the current observation for memory writing. Denote the output under geometry conditioning by $\mathcal{C}_\Theta^G(X_i, Z, P_i, m_{i-1})$. The recurrent update shows how geometry and confidence regulate the accumulated information:

$$m_t = \sum_{i=1}^{t} \omega(g, q_{i-1})\, \mathcal{C}_\Theta^G(X_i, Z, P_i, m_{i-1}),$$
$$\omega(g, q) = \sigma(\mathrm{MLP}([\widetilde{g}, q])), \tag{5}$$

where $\widetilde{g}$ denotes the normalized camera geometry and q denotes the tracking confidence. The gate learns how camera geometry and confidence affect the contribution of each observation to memory.

Following confidence-guided updates in STARK and MixFormer (Yan et al., 2021; Cui et al., 2022), GATP uses confidence and camera geometry to regulate memory updates. The decoded q_t controls the next update. During training, the memory is unrolled for $N = 3$ complete search frames with supervision at each frame. During inference, memory is updated throughout the sequence.

3.3 Geometry-conditioned decoding

GCD retains the baseline prediction head (Zhou et al., 2019; Ye et al., 2022) and augments the peak-plus-offset readout using neighboring responses and offsets conditioned on camera geometry. It uses response map $s_t \in [0,1]^{N_s \times N_s}$, offset map $o_t \in \mathbb{R}^{2 \times N_s \times N_s}$, and size map $d_t \in [0,1]^{2 \times N_s \times N_s}$, where N_s denotes the response map size.

Let $p^* = \arg\max_p s_t(p)$, and let $\mathcal{N}(p^*)$ denote the nine positions around this peak, where $p_i \in \mathcal{N}(p^*)$. The altitude-based mapping $\psi_{\mathrm{dec}}(h)$ provides the temperature $\tau_g > 0$ and relative-support threshold ρ_g. Define $r_i = s_t(p_i)/\max(s_t(p^*), \epsilon)$ and the relative-support gate $k_i = \sigma(\kappa_s(r_i - \rho_g))$. The continuous center is

$$(\hat{x}_t, \hat{y}_t) = \frac{\sum_i k_i e^{s_t(p_i)/\tau_g} \big(p_i + o_t(p_i)\big)}{N_s \sum_j k_j e^{s_t(p_j)/\tau_g}}, \tag{6}$$

where κ_s controls the sharpness of the sigmoid gate and ϵ avoids division by zero. Coordinates and offsets are normalized by N_s.

The box is $\hat{b}_t = \mathcal{M}_t([\hat{x}_t, \hat{y}_t, d_t(p^*)])$, where $\mathcal{M}_t$ maps crop coordinates to image coordinates. Confidence is obtained as $q_t = s_t(p^*)$.

The mapping ψ_{dec} uses fixed parameter pairs over altitude intervals, while pitch and focal length continue to condition upstream features. Although the decoder setting remains fixed within an interval, the aggregation weights adapt to each frame's response distribution. Thus, geometry adjusts the contribution of neighboring positions within a fixed neighborhood.

3.4 Distillation-Guided Geometry Estimation

For benchmarks without camera geometry annotations, we introduce Distillation-Guided Geometry Estimation to estimate camera geometry from visual features. The estimator is trained on GeoUAV, where measured camera geometry provides supervision following the learning using privileged information paradigm (Vapnik & Vashist, 2009). The estimator concatenates the separately mean-pooled early features from the template and search regions. It predicts camera geometry $(\hat{h}, \hat{\theta}, \hat{f})$ within the collection ranges.

After training the tracker with measured geometry, the tracker is frozen and the estimator is optimized on GeoUAV. The estimator uses ℓ_1 supervision on normalized geometry parameters and token-level ℓ_1 distillation. The frozen geometry encoder maps measured and estimated geometry to teacher and student tokens, which are aligned through generalized distillation (Lopez-Paz et al., 2016). At inference, camera geometry is estimated once from the initial template and the first search frame, then cached and reused by all geometry interfaces throughout the sequence.

4 The GeoUAV dataset

Data collection. We recorded the data with a DJI Mavic 4 Pro at 1280×720 resolution and 60 fps. The camera settings include altitudes of 30, 60, and 100 m, gimbal pitch angles of $30°$, $60°$, and $90°$, and focal lengths ranging from $1.0\times$ to $2.0\times$ with $0.1\times$ increments.

Annotation. Visible targets receive bounding boxes in each frame, and fully unobservable targets are marked absent. Each sequence includes a natural-language description of stable target attributes and relative position. The description remains consistent throughout the sequence to provide a stable reference for target identification. Camera settings remain fixed within each sequence and vary across sequences. Manual verification aligns target identity, descriptions, boxes, and camera records, ensuring that geometry annotations match the recorded camera settings. This partition ensures that frames from the same sequence do not appear in both sets.

Statistics and split. GeoUAV contains 495 sequences with 56K frames from 8 target classes. We divide the dataset into three subsets according to altitude: GeoUAV30, GeoUAV60, and GeoUAV100 correspond to sequences recorded at 30 m, 60 m, and 100 m, respectively. The sequences cover playgrounds, parks, parking lots, intersections, and urban arterials, with balanced distributions across camera settings. Figure 3 shows frequent small targets and an image distribution dominated by persons and cars. We partition complete sequences into 396 training and 99 test sequences.

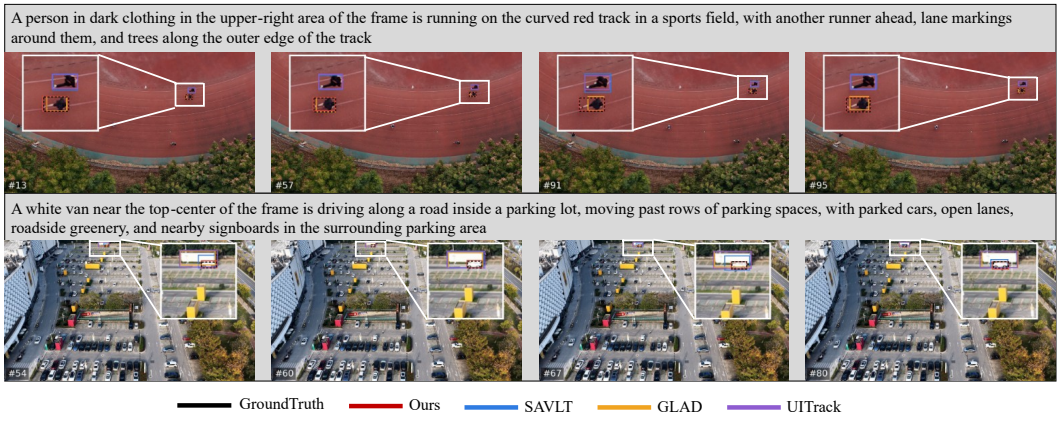


Figure 4: Qualitative comparison on GeoUAV.

Position among existing datasets. Table 1 compares GeoUAV with existing UAV tracking datasets. Most datasets provide aerial videos and tracking annotations without language descriptions or numerical camera geometry annotations. UAVDT provides coarse altitude and viewpoint categories, but does not contain numerical camera geometry annotations. WebUAV-3M provides large-scale UAV tracking data with natural-language descriptions, but does not include numerical camera geometry annotations. GeoUAV provides natural language descriptions and numerical camera geometry annotations under controlled camera settings.

Table 1: Comparison with aerial tracking datasets. H, P, and F denote altitude, gimbal pitch angle, and focal length annotations.

Dataset	Absent	Lang.	H	P	F
UAV123 (Mueller et al., 2016)	✗	✗	✗	✗	✗
UAV20L (Mueller et al., 2016)	✗	✗	✗	✗	✗
DTB70 (Li & Yeung, 2017)	✗	✗	✗	✗	✗
UAVDT (Du et al., 2018)	✗	✗	✓	✓	✗
UAVDark135 (Li et al., 2021)	✗	✗	✗	✗	✗
VisDrone-SOT2019 (Du et al., 2019)	✗	✗	✗	✗	✗
WebUAV-3M (Zhang et al., 2023)	✓	✓	✗	✗	✗
GeoUAV (Ours)	✓	✓	✓	✓	✓

Table 2: Results on GeoUAV30, GeoUAV60, and GeoUAV100.

Method	GeoUAV30			GeoUAV60			GeoUAV100		
	AUC	P	P_n	AUC	P	P_n	AUC	P	P_n
Vision-only									
ORTrack (Wu et al., 2025)	73.83	91.22	89.06	55.22	74.97	63.41	51.31	77.59	58.32
AVTrack (Li et al., 2024)	77.07	91.86	91.86	61.80	80.98	72.12	54.73	81.68	62.77
Vision–language									
UITrack (Wang et al., 2025)	79.66	95.80	93.15	67.08	87.53	79.24	65.46	90.32	75.71
ATCTrack (Feng et al., 2025)	79.24	94.49	91.87	70.15	87.27	81.40	67.31	91.13	76.87
SAVLT (Zhao et al., 2025)	80.00	95.57	93.12	70.72	89.29	81.23	65.86	89.59	74.98
GLAD (Luo et al., 2026)	67.84	86.63	82.70	56.30	83.00	65.15	47.91	74.15	54.75
GeoVLT (Ours)	**82.55**	**96.81**	**94.84**	**74.58**	**92.27**	**87.81**	**69.69**	**93.15**	**78.71**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Training and evaluation. We train GeoVLT on the GeoUAV training split for 40 epochs using AdamW. The batch size is 8 and the initial learning rate is 10^{-4}. Altitude groups are sampled with equal probability. All experiments use one NVIDIA GeForce RTX 5090. GATP uses a temporal unroll of $N = 3$, and GCD uses a neighborhood radius of one with height-conditioned readout parameters. Evaluation reports success AUC, center precision (P), and normalized precision (P_n), with macro scores computed as equally weighted means across altitude groups.

5.2 COMPARISON WITH STATE OF THE ART

Overall comparison. Table 2 evaluates tracking performance under different flight settings and compares vision-only and vision-language trackers. Compared with vision-only trackers, GeoVLT achieves AUC improvements of 5.48%, 12.78%, and 14.96% on GeoUAV30, GeoUAV60, and GeoUAV100, respectively. These improvements show that camera geometry provides additional cues beyond visual appearance, especially when altitude changes cause larger variations in target scale and visible detail. Compared with vision-language trackers, GeoVLT further improves AUC by 2.55%, 3.86%, and 2.38% on the three altitude groups. Although language descriptions provide semantic target information, they do not directly describe the imaging conditions caused by camera geometry. By incorporating camera geometry, GeoVLT maintains consistent improvements across different flight settings and achieves the highest AUC, P, and P_n on all three subsets.

Table 3: Progressive ablation on GeoUAV30, GeoUAV60, and GeoUAV100.

Setting			GeoUAV30			GeoUAV60			GeoUAV100		
GCR	GATP	GCD	AUC	P	P_n	AUC	P	P_n	AUC	P	P_n
			79.71	94.14	92.00	71.17	90.29	81.84	65.62	89.40	74.25
✓			81.91	95.05	94.27	73.01	90.92	85.10	66.71	89.65	76.12
✓	✓		82.07	96.04	94.40	73.48	91.35	86.15	69.53	93.07	78.20
✓	✓	✓	**82.55**	**96.81**	**94.84**	**74.58**	**92.27**	**87.81**	**69.69**	**93.15**	**78.71**

Table 4: **Validity of camera geometry inputs.** Recorded geometry is compared with a fixed tuple shared across sequences and geometry tuples taken from other sequences.

Input	GeoUAV30			GeoUAV60			GeoUAV100		
	AUC	P	P_n	AUC	P	P_n	AUC	P	P_n
Correct	**82.55**	**96.81**	94.84	**74.58**	**92.27**	**87.81**	**69.69**	**93.15**	**78.71**
Fixed	82.35	96.8	**95.11**	73.43	91.47	86.11	66.14	88.06	74.84
Shuffled	81.01	94.84	94.12	73.08	90.02	86.34	67.80	90.64	76.08

Visualization analysis. To further analyze the tracking behavior of different methods, we visualize tracking results and response maps in Figures 4 and 5. In the running sequence, similar objects near the target cause competing trackers to produce responses around both objects, while GeoVLT focuses on the annotated target. For the second sequence with limited target detail, other trackers generate scattered responses, whereas GeoVLT maintains a clear response around the target region. In the third sequence, where the target has similar appearance to the background, competing trackers respond to a larger surrounding area, while GeoVLT produces a more concentrated response. These results indicate that camera geometry helps the tracker interpret target appearance changes and maintain reliable tracking under different imaging conditions.

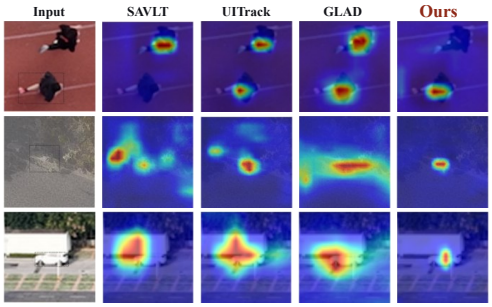


Figure 5: Response map comparison under different tracking conditions.

5.3 Ablation and Analysis

Component ablation. Table 3 evaluates the contribution of GCR, GATP, and GCD by progressively adding each component. Adding all three modules improves AUC across GeoUAV30, GeoUAV60, and GeoUAV100, showing the effect of camera geometry under different flight settings. GCR provides the largest gain on GeoUAV30, improving AUC by 2.20%. This indicates that GCR helps interpret target appearance in the current frame. After adding GATP, the largest improvement appears on GeoUAV100, with a 2.82% AUC gain, where target detail is limited and historical information provides additional cues for maintaining target identity. Adding GCD further improves GeoUAV60 by 1.10%, showing that GCD helps refine predictions under different imaging conditions.

Camera geometry input validity. To evaluate the effectiveness of camera geometry information, we compare three types of geometry inputs: Correct uses the recorded geometry of each sequence, Fixed uses a constant geometry tuple for all sequences, and Shuffled uses geometry from another sequence. Table 4 shows that correct geometry achieves the best AUC across all altitude groups. Compared with correct geometry, fixed geometry reduces AUC by 3.55% on GeoUAV100, where target detail is limited. Shuffled geometry also decreases AUC from 82.55% to 81.01% on GeoUAV30, showing that geometry inconsistent with the current sequence cannot provide the same tracking cues. The comparable P under fixed geometry on GeoUAV30 may be attributed to richer target detail at lower altitude, where small prediction deviations are less sensitive to geometry mismatch and can still

Table 5: Generalization to other UAV datasets using estimated camera geometry under vision–language and vision-only protocols.

Method	Vision–language				Vision-only			
	UAV20L-NLP		DTB70-NLP		UAV123		DTB70	
	AUC	P	AUC	P	AUC	P	AUC	P
SAVLT	59.39	75.71	54.41	71.92	59.43	80.82	52.60	69.29
UITrack	61.66	77.65	51.33	57.22	56.88	77.30	50.20	65.65
GLAD	35.14	46.90	39.28	54.21	40.03	58.67	36.68	52.28
GeoVLT (Ours)	**64.11**	**80.64**	**54.80**	**72.36**	**62.07**	**83.54**	**54.91**	**73.31**

satisfy the fixed pixel threshold of P. These results demonstrate that camera geometry is useful when it corresponds to the current imaging conditions.

Generalization across datasets. To improve the generalization ability of GeoVLT, we apply the geometry estimator to public UAV benchmarks without camera geometry annotations. The estimated geometry is obtained from the first frame and used without adaptation to the test dataset. Under the vision–language protocol, the NLP benchmarks use the language annotations provided by AVLTrack (Xue et al., 2025). Table 5 shows that GeoVLT maintains the best performance under both vision–language and vision-only protocols. In the vision-only setting, GeoVLT improves AUC by 2.64% on UAV123 and 2.31% on DTB70 compared with the best competing trackers. Under the vision–language setting, GeoVLT also achieves consistent gains, including a 2.45% AUC improvement on UAV20L-NLP. These results indicate that estimated camera geometry remains useful when the test dataset does not provide geometry annotations and language input is unavailable.

5.4 Hyperparameter sensitivity

Temporal unroll length. The temporal unroll length N determines the number of consecutive frames used to train the recurrent memory, while inference updates the memory throughout the sequence. Figure 6 in Appendix B shows that AUC improves as N increases and reaches the highest value at $N = 3$ across all altitude groups. Although $N = 2$ obtains slightly higher P, $N = 3$ provides better AUC and P_n, indicating a better balance between center accuracy and overall tracking quality. Increasing N to four or five frames reduces AUC, showing that longer training unrolls do not further improve memory propagation. Therefore, $N = 3$ is selected as the default setting.

GCD temperature and relative response threshold. To analyze the effect of the two parameters in GCD, we evaluate different combinations of temperature τ and relative-support threshold ρ under three altitude settings. Figures 7–9 in Appendix B show that the selected parameter pair varies with altitude. The retained pairs are $(0.15, 0.70)$ for GeoUAV30, $(0.10, 0.70)$ for GeoUAV60, and $(0.10, 0.50)$ for GeoUAV100. The parameter sensitivity differs across altitude groups. On GeoUAV100, changing the parameter combination produces larger performance variations, while GeoUAV30 remains relatively stable. This indicates that response aggregation is more sensitive when target detail is limited. The selected pairs achieve a balance among AUC, P, and P_n for each setting.

6 Conclusion

We study UAV vision–language tracking under different camera settings, where changes in camera geometry affect target appearance and scene coverage. We construct GeoUAV with measured camera geometry and introduce GeoVLT to incorporate this information into target representation, temporal memory, and coordinate prediction. Experiments show that tracking benefits from geometry matched to the current imaging conditions, while geometry estimated from the first frame supports tracking on public UAV benchmarks without geometry annotations. Extending it to camera settings that change during tracking remains a direction for future work.

REFERENCES

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025.

Wenrui Cai, Qingjie Liu, and Yunhong Wang. HIPTrack: Visual tracking with historical prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19258–19267, 2024.

Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, and Changhong Fu. TCTrack: Temporal contexts for aerial tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14798–14808, 2022.

Chenglizhao Chen, Shaofeng Liang, Runwei Guan, Xiaolou Sun, Haocheng Zhao, Haiyun Jiang, Tao Huang, Henghui Ding, and Qing-Long Han. AerialMind: Towards referring multi-object tracking in UAV scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2805–2813, 2026a.

Shih-Fang Chen, Jun-Cheng Chen, I-Hong Jhuo, and Yen-Yu Lin. GOT-Edit: Geometry-aware generic object tracking via online model editing. In *International Conference on Learning Representations*, 2026b.

Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. MixFormer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13608–13618, 2022.

Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision*, pp. 370–386, 2018.

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 213–226, 2019.

Jose M. Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11826–11835, 2019.

Xiaokun Feng, Shiyu Hu, Xuchen Li, Dailing Zhang, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. ATCTrack: Aligning target-context cues with dynamic target states for robust vision-language tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19850–19861, 2025.

Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rares Ambrus, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9233–9243, 2023.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1501–1510, 2017.

Grzegorz Jacenków, Alison Q. O'Neil, Brian Mohr, and Sotirios A. Tsaftaris. INSIDE: Steering spatial attention with non-imaging information in CNNs. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 385–395, 2020.

Bowen Li, Changhong Fu, Fangqiang Ding, Junjie Ye, and Fuling Lin. All-day object tracking for unmanned aerial vehicle. *arXiv preprint arXiv:2101.08446*, 2021.

Siyi Li and Dit-Yan Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 4140–4146, 2017.

Xiaohai Li, Bineng Zhong, Qihua Liang, Zhiyi Mo, Jian Nong, and Shuxiang Song. Dynamic updates for language adaptation in visual-language tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19165–19174, 2025.

Yongxin Li, Mengyuan Liu, You Wu, Xucheng Wang, Xiangyang Yang, and Shuiwang Li. Learning adaptive and view-invariant vision transformer for real-time UAV tracking. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 28403–28420, 2024.

Zhenyang Li, Ran Tao, Efstratios Gavves, Cees G. M. Snoek, and Arnold W. M. Smeulders. Tracking by natural language specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6495–6503, 2017.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016.

Xingyu Luo, Yidong Cai, Jie Liu, Jie Tang, Gangshan Wu, and Limin Wang. GLAD: Generative language-assisted visual tracking for low-semantic templates. *arXiv preprint arXiv:2602.00570*, 2026.

Yinchao Ma, Yuyang Tang, Wenfei Yang, Tianzhu Zhang, Jinpeng Zhang, and Mengxue Kang. Unifying visual and vision-language tracking via contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4107–4116, 2024.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, pp. 405–421, 2020.

Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for UAV tracking. In *Proceedings of the European Conference on Computer Vision*, pp. 445–461, 2016.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019.

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, pp. 3942–3951, 2018.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

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 International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision*, pp. 529–545, 2018.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.

Jingchao Wang, Zhijian Wu, Wenlong Zhang, Wenhui Liu, Jianwei Zhang, and Dingjiang Huang. Overcoming feature contamination by unidirectional information modeling for vision-language tracking. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2025.

Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1571–1580, 2021a.

Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13763–13773, 2021b.

Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 606–615, 2018.

You Wu, Xucheng Wang, Xiangyang Yang, Mengyuan Liu, Dan Zeng, Hengzhou Ye, and Shuiwang Li. Learning occlusion-robust vision transformers for real-time UAV tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17103–17113, 2025.

Zhenyu Wu, Karthik Suresh, Priya Narayanan, Hongyu Xu, Heesung Kwon, and Zhangyang Wang. Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1201–1210, 2019.

Fei Xie, Lei Chu, Jiahao Li, Yan Lu, and Chao Ma. VideoTrack: Learning to track objects via video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22826–22835, 2023.

Yuanliang Xue, Bineng Zhong, Guodong Jin, Tao Shen, Lining Tan, Ning Li, and Yaozong Zheng. AVLTrack: Dynamic sparse learning for aerial vision-language tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(8):7554–7567, 2025.

Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10448–10457, 2021.

Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Proceedings of the European Conference on Computer Vision*, pp. 341–357, 2022.

Kai Ye, Haidi Tang, Bowen Liu, Pingyang Dai, Liujuan Cao, and Rongrong Ji. More clear, more flexible, more precise: A comprehensive oriented object detection benchmark for UAV, 2025.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3D prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9043–9053, 2023.

Chunhui Zhang, Guanjie Huang, Li Liu, Shan Huang, Yinan Yang, Xiang Wan, Shiming Ge, and Dacheng Tao. WebUAV-3M: A benchmark for unveiling the power of million-scale deep UAV tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9186–9205, 2023.

Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7091–7100, 2020.

Jie Zhao, Xin Chen, Shengming Li, Chunjuan Bo, Dong Wang, and Huchuan Lu. Self-adaptive vision-language tracking with context prompting. *IEEE Transactions on Image Processing*, 34: 8046–8058, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16795–16804, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022b.

Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23151–23160, 2023.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

A LLM USAGE

Large language models were used to assist with drafting and polishing the language of this manuscript, including sentence rephrasing, grammar correction, and improving clarity. The authors developed the research ideas, methodology, analyses, and conclusions, reviewed all AI-assisted text, and take full responsibility for the manuscript's content.

B ADDITIONAL HYPERPARAMETER ANALYSIS

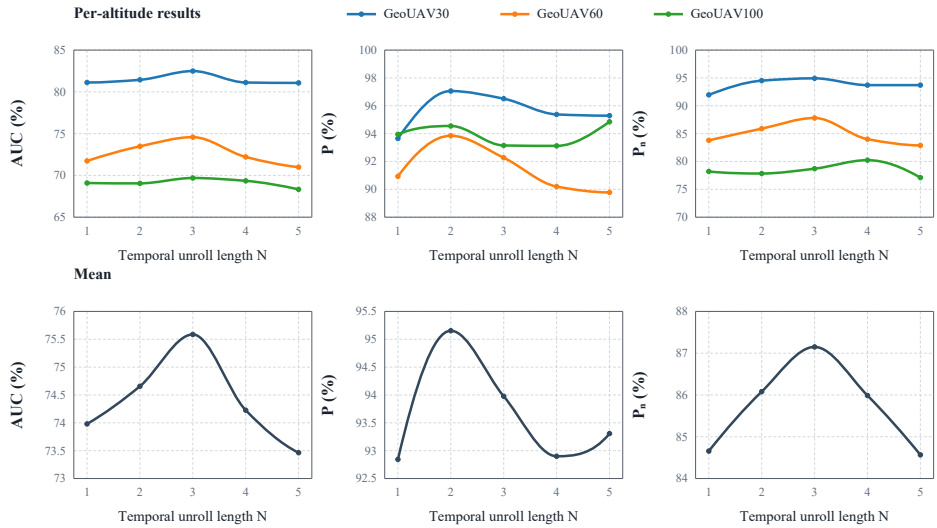


Figure 6: Effect of temporal unroll length N. The upper row reports each altitude group, and the lower row shows the mean across the three groups.

B.1 TEMPORAL UNROLL LENGTH

The temporal unroll length N determines how many continuous frames are used to train GATP's recurrent memory. To examine the effect of training history, this experiment varies N from one to five while keeping the other parameters fixed. Figure 6 reports the results for each altitude group and the mean across the three groups.

Temporal horizon. Figure 6 shows that $N = 3$ achieves the highest AUC in all three altitude groups. The mean AUC also reaches its maximum of 75.59% at this setting. Increasing N beyond three reduces AUC across all groups. This suggests that additional training history can improve memory learning, but a longer unroll does not guarantee better tracking.

The precision results reveal a different pattern. The highest mean P of 95.15% occurs at $N = 2$, while $N = 3$ achieves the highest mean P_n of 87.15%. GeoUAV100 also reaches its highest P_n at $N = 4$. The default is therefore set to $N = 3$ based on its AUC across all altitude groups and its mean normalized precision. This choice accepts a reduction in P compared with $N = 2$.

Training and inference. The sweep changes the number of frames used to train the recurrent memory. During inference, memory continues to update throughout the sequence. Thus, N controls the training unroll and does not limit the duration of the available tracking history.

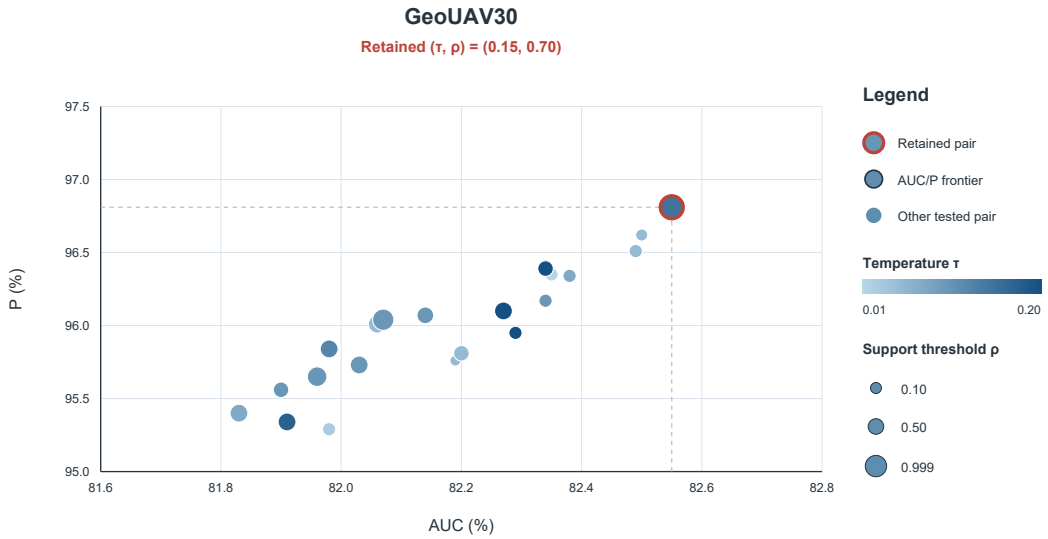


Figure 7: GCD parameter sensitivity on GeoUAV30. The retained pair $(\tau, \rho) = (0.15, 0.70)$ achieves the highest measured AUC and P.

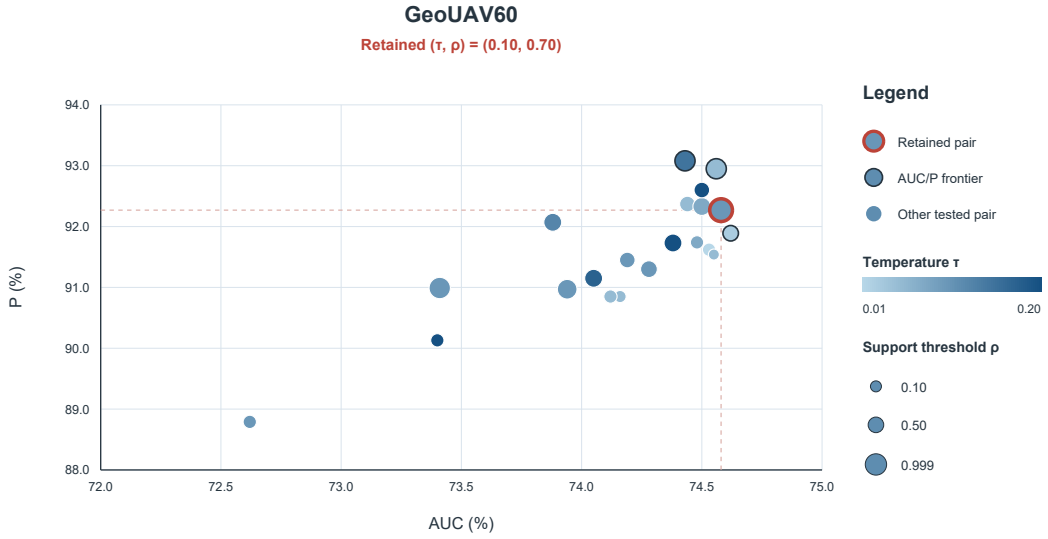


Figure 8: GCD parameter sensitivity on GeoUAV60. The retained pair $(\tau, \rho) = (0.10, 0.70)$ lies on the AUC/P frontier.

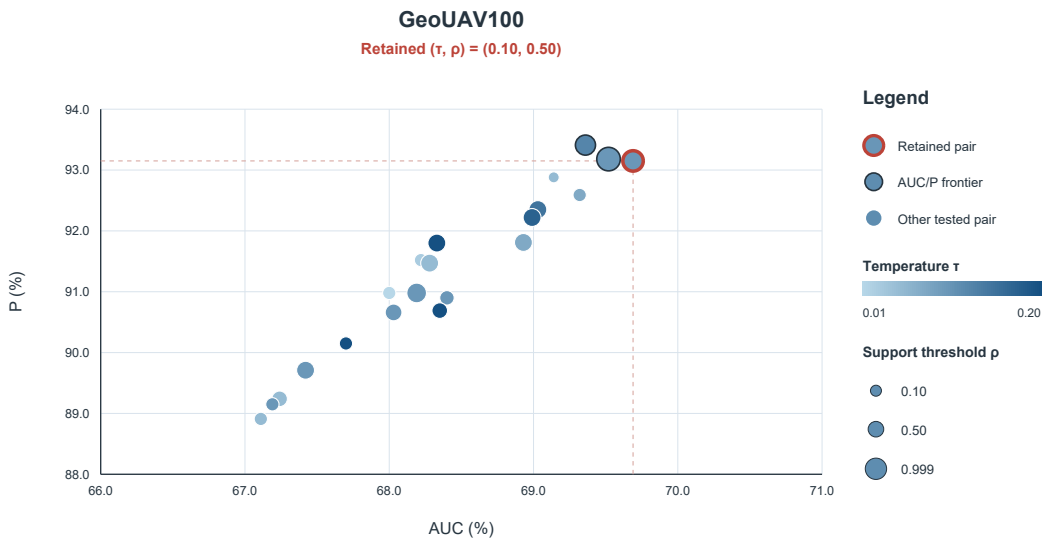


Figure 9: GCD parameter sensitivity on GeoUAV100. The retained pair $(\tau, \rho) = (0.10, 0.50)$ achieves the highest measured AUC. Settings with higher P yield lower AUC.

B.2 GCD TEMPERATURE AND RELATIVE SUPPORT

To examine the effect of GCD's parameters, this experiment evaluates 22 combinations of temperature τ and relative-support threshold ρ in each altitude group. Temperature controls the concentration of response weights, while the threshold regulates the contribution of weaker neighboring responses. Both parameters affect the final coordinate prediction.

Figures 7–9 show the relationship between AUC and P for the evaluated combinations. Each marker represents one parameter pair. Color indicates τ, and marker size indicates ρ. The red ring identifies the retained setting.

GeoUAV30. Figure 7 shows a relatively narrow AUC distribution across the evaluated pairs. The retained setting $(0.15, 0.70)$ achieves the highest AUC of 82.55% and P of 96.81%. Unlike the other altitude groups, the same pair maximizes both displayed metrics. The limited variation in AUC also suggests that coordinate prediction is relatively stable across the tested parameters at this altitude.

GeoUAV60. Figure 8 shows that AUC and P favor different parameter combinations. The highest AUC of 74.62% occurs at $(0.025, 0.30)$, where P is 91.89%. The retained pair $(0.10, 0.70)$ achieves a similar AUC of 74.58% and increases P to 92.27%. Other pairs obtain higher P but produce lower AUC. The retained setting therefore offers a measured trade-off between the two metrics, with only a small reduction from the highest AUC.

GeoUAV100. Figure 9 shows a wider AUC distribution than GeoUAV30. This indicates greater sensitivity to the tested readout parameters at higher altitude. The retained pair $(0.10, 0.50)$ achieves the highest AUC of 69.69% with a P of 93.15%. Increasing P to 93.41% at $(0.125, 0.70)$ reduces AUC to 69.36%. The retained setting prioritizes AUC while keeping P close to its maximum.

Parameter selection. The three altitude groups exhibit different relationships between AUC and P. GeoUAV30 has a single tested pair that maximizes both metrics. GeoUAV60 requires a trade-off between them, while GeoUAV100 shows greater variation in AUC across the evaluated parameters. These differences motivate the use of separate parameter pairs for the three altitude groups.

The selected parameters remain fixed within each sequence because its camera geometry does not change. However, the weights assigned to neighboring responses still vary with the current frame. The reported results support the retained settings among the tested combinations but do not establish an optimum over the full parameter space.