Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

VQATrack:Visual Question Answering Model-Driven Lightweight UAV Visual-Language Object Tracking

Lingling Yang
Xi'an Technological University
Liqiang Liu
liuliqiang@xatu.edu.cn
Xi'an Technological University
Yanfang Fu
Xi'an Technological University
Anyuan Xie
Ministry of Natural Resources
Yujia Li
Xi'an Technological University
Ningchao Zhang
Xi'an Technological University

Research Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11561893/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

VQATrack:Visual Question Answering Model-Driven Lightweight UAV Visual-Language Object Tracking

Lingling Yang[1,3], Liqiang Liu[1*], Yanfang Fu[1*], Anyuan Xie[2], Yujia Li[3], Ningchao Zhang[3]

[1*]School of Computer Science and Engineering, Xi'an Technological University, Xi'an, 710021, P.R.China.
[2]Key Laboratory of Marine Environmental Survey Technology and Application, Ministry of Natural Resources, Organization, Street, Guangzhou, 510000, P.R.China.
[3]School of Electronic Information Engineering, Xi'an Technological University, Xi'an, 710021, P.R.China.

*Corresponding author(s). E-mail(s): liuliqiang@xatu.edu.cn; fuyanfang@xatu.edu.cn;
Contributing authors: yanglingling1214@163.com; xay_smst@163.com; liyyujia@163.com; zhangningchao@xatu.edu.cn;

Abstract

Although multi-modal target tracking has achieved promising performance in addressing dynamic targets and background clutter for UAV tracking in complex scenarios, existing methods still suffer from two critical bottlenecks: limited semantic information from manual language annotations and challenges in real-time onboard deployment. Furthermore, most existing studies are restricted to dataset-level evaluation and lack systematic closed-loop validation. To tackle these issues, this paper proposes VQATrack, a lightweight vision-language multi-modal tracking and control framework for autonomous UAV tracking. First, to alleviate the inefficiency of manual language description, we design a visual question answering-based language generation module, which automatically translates visual observations into structured semantic descriptions. Second, we construct a lightweight vision-language tracking model: computational redundancy is reduced through hierarchical gradient freezing, background interference is suppressed via visual token masking, and efficient interaction

between visual and language features is achieved by hierarchical dynamic fusion. Finally, we build the first closed-loop simulation platform that integrates target perception, depth estimation, obstacle avoidance, and motion control for comprehensive validation of UAV deployment. Extensive experiments verify the effectiveness and superiority of VQATrack, which achieves 66.2% accuracy on the TNL2K dataset and 62 FPS on the RTX 3090.The code is available at:https://github.com/yll1214/VQATrack.

Keywords: Text Generation; Cross-Modal Interaction; Target Tracking

1 Introduction

UAV technology is now widely used in aerial inspection and surveillance. Visual object tracking, a core task for autonomous flight, enables UAVs to lock onto and follow targets in dynamic environments [1], but faces challenges from real-world complexity and uncertainty [2].

As shown in Figure 1(a), early tracking methods relied on single-modality visual information using correlation filtering or siamese networks [3–6]. These methods struggle with appearance changes or distractors, especially in UAV views where illumination changes and jitter amplify instability [7].

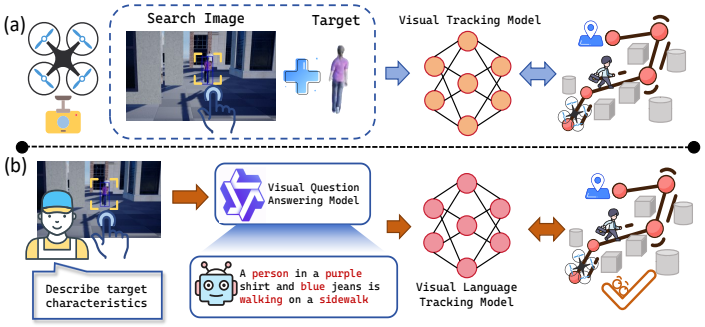


Fig. 1 (a) UAV achieves target tracking through correlation filtering or siamese networks. (b) Vision-language target tracking for UAVs driven by a visual question answering model.

Vision-language multi-modal tracking offers a solution [8], but no UAV-specific research exists. Deployment bottlenecks include: 1) Unrealistic reliance on manual annotations; 2) High computational cost on embedded platforms [9–11]; 3) Dataset-only evaluation ignoring real-world factors [12].To address these challenges, as shown in Figure 1(b), we propose VQATrack, a lightweight vision-language multi-modal tracking and control method for autonomous UAV tracking, and construct a full-link closed-loop tracking simulation platform for practical UAV deployment. Specifically, our main contributions include:

- A visual question answering-based module automatically generates structured target descriptions from visual input, with multi-image strategy mitigating hallucination.
- A lightweight vision-language tracking model accelerates inference through gradient freezing, token masking, and hierarchical dynamic fusion for efficient cross-modal interaction.
- The first simulation-based closed-loop platform integrating target description, perception, depth estimation, obstacle avoidance, and control, enabling stable open-world tracking.

2 Related Work

2.1 Single-Modality Object Tracking Algorithms for UAVs

Single-modality object tracking for UAVs has shown broad applications in aerial video analysis and intelligent surveillance, yet faces challenges including viewpoint variations, multi-scale changes, and small target capture. Researchers have explored various solutions. Cao et al. Cao et al. [13] constructed a self-attention and cross-attention aggregation network for complex semantic variations. Yao et al. Yao et al. [14] introduced a saliency-guided mechanism for foreground-background discrimination in Transformers. Wu et al. Wu et al. [15] designed an adaptive module for viewpoint-invariant features via mutual information maximization. Benchmark datasets including UAV123 [16], UAVDark135 [17], and NAT2024 [18] support algorithm evaluation in air-to-ground scenarios.

For systematic robustness evaluation, Zhao et al. [19] constructed BioDrone, the first bionic flapping-wing UAV tracking benchmark. Wu et al. [20] proposed ORTrack, an occlusion-robust method with spatial Cox processes and knowledge distillation. Chen et al. [21] proposed HETrack for enhanced small target perception. Fu et al. [2] explored progressive representation learning, while Wen et al. [1] proposed SDGTrack for resource scheduling. However, tracking stability and system collaboration efficiency still require further validation.

2.2 Multi-Modal Object Tracking Algorithms for UAVs

Research in vision-language object tracking for UAVs remains limited, primarily due to two factors: dataset scarcity and domain gap. While general vision-language tracking benchmarks like TNL2K Wang et al. [22], OTB99 Wu et al. [23], LASOT [24], and RefCOCOg [25] are well-established, they are collected from ground or fixed viewpoints, lacking UAV-specific characteristics such as top-down views and dynamic aerial footage. Recent efforts have addressed this gap. Zhang et al. [26] constructed the WebUAV-3M dataset with natural language and audio descriptions, establishing the UTUSC evaluation protocol and seven challenging sub-test sets, advancing multimodal UAV tracking research. Xue et al. [27] augmented DTB70 and UAV20L with text descriptions, introducing natural language into aerial tracking. Their dynamic sparse learning (DSL) module and multi-level language perception (MLP) strategy achieved robust performance under complex backgrounds and target deformation.

Furthermore, to promote the development of multi-modal object tracking for UAVs, the large-scale general visible-thermal infrared UAV tracking benchmark established by Zhang et al. [28] established a large-scale visible-thermal infrared UAV tracking benchmark for fair evaluation. At the algorithmic level, Xue et al. [29] proposed SiamCAF, a complementary attention fusion tracker with a coupled feature fusion (CCF) module for effective visible-thermal information integration. Gao et al. [30] introduced a Transformer-based dual-modal feature aggregation network that adaptively fuses RGB-T features through spatial and channel attention mechanisms, achieving lightweight efficiency and robust performance under complex illumination and occlusion.

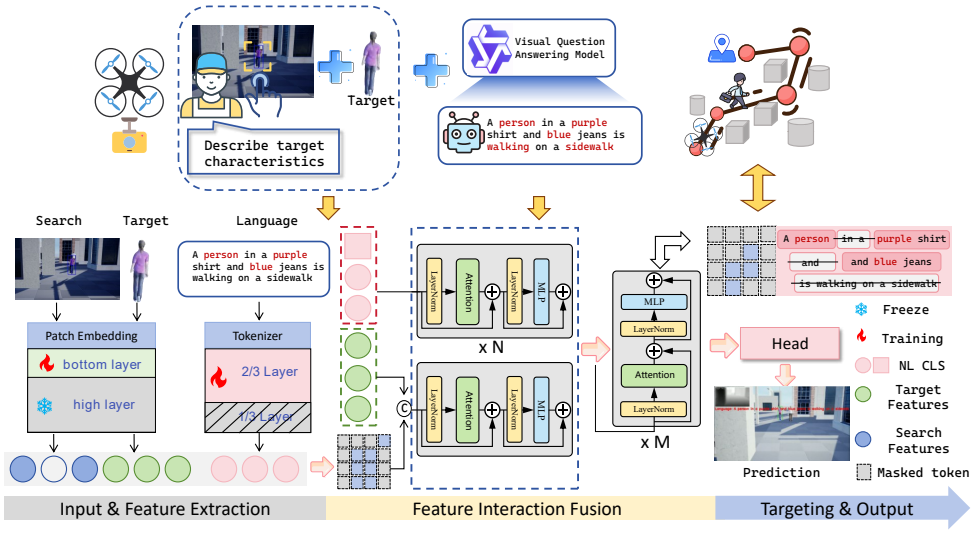


Fig. 2 Overview of our proposed method. A visual question answering-based language generation module is employed to automatically convert visual inputs into structured descriptions. A lightweight vision-language tracking model is constructed using a hierarchical gradient freezing strategy to reduce computational redundancy. A visual token masking mechanism is introduced to suppress background interference, and hierarchical dynamic fusion is adopted to achieve efficient interaction between visual and language features.

3 Methods

In this paper, we propose VQATrack, a lightweight vision-language multi-modal tracking framework designed for UAV embedded platforms. As illustrated in Figure 2, this method aims to achieve a full-link closed loop from target semantic understanding to autonomous flight control.

First, the target language description generation module, based on the visual question answering architecture Qwen3-VL-4B-Instruct-FP8[31], automatically converts target images captured from UAV perspectives into structured natural language

descriptions, eliminating reliance on manual annotations. Second, the lightweight vision-language tracking model receives language descriptions and search images as inputs, reduces feature encoding computational redundancy through a hierarchical gradient freezing strategy, suppresses background interference with a visual token masking mechanism, and achieves efficient interaction between visual and language features via a hierarchical dynamic fusion mechanism, ultimately outputting target bounding boxes. Finally, the UAV autonomous tracking module integrates depth information and achieves stable target following through a depth map-based obstacle avoidance strategy and velocity control commands. Ultimately, VQATrack forms a complete closed loop encompassing visual perception, semantic understanding, target localization, and flight control, enabling UAVs to achieve autonomous target tracking and real-time obstacle avoidance in mapless open-world environments.

3.1 Target Language Description Generation

Given the inefficiency and coarse semantic granularity of manual language descriptions in real-world UAV tracking tasks, we propose a target semantic description generation method based on a visual question answering (VQA) architecture. VQATrack employs Qwen3-VL-4B-Instruct-FP8[31] as its core generation model, achieving fast and high-quality conversion from visual input to structured language descriptions through a multi-stage cross-modal alignment mechanism. Qwen3-VL-4B-Instruct-FP8 reduces GPU memory usage by 40% while maintaining performance nearly identical to the original FP16 model, enabling smooth operation on consumer-grade graphics cards with 8GB of memory. This model performs joint modeling of images and text within the Transformer framework, learning sequence generation patterns under conditions where both textual and visual information are jointly present:

$$\mathcal{L} = \sum_{t=1}^{T} \log P_\theta(y_t | y_{<t}, x_{\text{text}}, x_{\text{image}}) \tag{1}$$

where $\mathcal{L}$ is the objective function for model training, T is the length of the output sequence, y_t is the t-th generated token, $y_{<t}$ denotes the sequence of tokens generated before position t, x_{text} is the input text sequence, x_{image} is the visual representation of the input image, and $P_\theta(\cdot)$ represents the conditional probability distribution parameterized by the model parameters θ.

Subsequently, we found that literature [32] points out that a single image input may cause model hallucination, generating non-existent objects or descriptions. Therefore, as shown in Figure 3, we discovered that diversified target image inputs can effectively reduce semantic discrepancies in language description results. We enable interactive selection of target objects from the UAV perspective, constructing structured semantic templates. These templates, together with current video frames, are input into the vision-language model to generate task-aligned natural language descriptions, achieving automatic conversion from structured attributes to natural language expressions.

Regarding the impact of input resolution on semantic consistency, we set template and search image resolutions to no less than 300×300, performing proportional

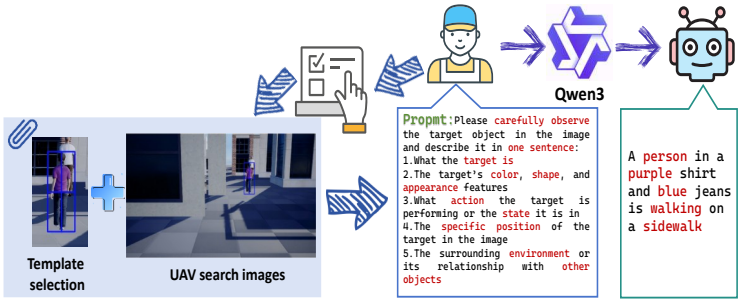


Fig. 3 Target text description generation method based on visual question answering model

upsampling. Experimental results show that below this threshold, generated descriptions exhibit significant fine-grained semantic deviations, validating the necessity of sufficient visual information for cross-modal semantic alignment.

3.2 Vision-Language Tracking Model

To adapt to UAV embedded devices, VQATrack designs a lightweight object tracking model and further optimizes it to meet the requirements of our task. As shown in Figure 2, our object tracking method supports two input modalities: natural language and UAV visual data. The tracking process with natural language descriptions in our framework can be summarized as follows: 1) In the first frame, we perform visual grounding using natural language, zero-padding tokens, and the search image as inputs, obtaining the target template based on the grounding results. 2) In each subsequent frame, we perform language-assisted visual tracking using natural language, the target template, and the search image.

3.2.1 Vision-Language Feature Encoding.

In multi-modal tracking, the efficiency and quality of visual and language feature extraction directly impact performance. We employ pre-trained BERT[33] and MAE[34] to encode natural language and search images, respectively. Based on findings in [35] that larger language models slightly mitigate text bias, and to balance capacity with computational overhead on UAV platforms, we design a hierarchical feature extraction and gradient freezing strategy.

For the language branch, we observe that the full 12-layer BERT exhibits computational redundancy in tracking tasks. Since text descriptions primarily provide target semantic priors rather than requiring complex language understanding, deep language modeling contributes limited improvement. Therefore, we remove the last 4 layers while retaining the first 8 layers, reducing computational burden while preserving semantic representation capabilities.

For the visual branch, the input image $I \in \mathbb{R}^{H\times W\times 3}$ is transformed into a sequence of visual tokens through linear projection. Considering the distinct roles of features at different layers in tracking tasks, we design a hierarchical freezing strategy. Low-layer features primarily encode general visual information such as edges and textures,

which have been sufficiently learned during large-scale pre-training and require minimal adjustment for tracking tasks. Therefore, we freeze the first $M = 3$ layers, with their gradients satisfying:

$$\frac{\partial \mathcal{L}}{\partial \theta_l} = 0, \quad \forall l \leq M \tag{2}$$

where θ_l represents the trainable parameters of the l-th layer, and $\mathcal{L}$ denotes the tracking loss function.

For high-layer features ($l > M$), which focus more on semantic information and are closely related to specific tracking targets, we keep them trainable to adapt to appearance variations of particular targets. Additionally, we introduce a loss-based dynamic unfreezing mechanism. Let $\mathcal{L}_{\text{val}}^k$ denote the validation loss at the k-th training epoch. If for $P = 3$ consecutive epochs, the following condition holds:

$$\mathcal{L}_{\text{val}}^k - \mathcal{L}_{\text{val}}^{k-P} < \delta = 0.01 \tag{3}$$

then the next layer is unfrozen, allowing more parameters to participate in training to enhance model adaptability.

3.2.2 Search Image Visual Token Masking Strategy.

The feature-encoded image contains rich semantic information. In visual object tracking tasks, effectively suppressing background interference is crucial for improving tracking robustness. To this end, inspired by the insights from [36], we introduce a background suppression mechanism based on self-similarity measurement during the feature extraction stage. By dynamically identifying background regions in the search image and attenuating their feature responses, the model is enabled to focus on the target foreground.

Given the search image x, it is first converted into a sequence of feature tokens through a patch embedding layer: $X \in \mathbb{R}^{B \times N_x \times D}$. We compute the maximum cosine similarity between each token and all preceding tokens within the same image:

$$s_{b,i} = \max_j \frac{X_{b,i} \cdot X_{b,j}}{\|X_{b,i}\| \|X_{b,j}\|} \tag{4}$$

where $s_{b,i}$ denotes the background confidence of the i-th token in the b-th sample, and $X_{b,i}$ and $X_{b,j}$ are the feature vectors of the i-th and j-th tokens, respectively.

Since background regions typically contain repetitive textures, they exhibit high similarity between tokens. Therefore, when the number of search image tokens is $N_x = 256$, we set an upper threshold $\tau = 0.8$, and tokens with similarity exceeding τ are identified as background. A binary mask matrix $M \in \{0, 1\}^{B \times N_x}$ is constructed as:

$$M_{b,i} = \mathbb{1}(s_{b,i} > \tau) \tag{5}$$

where $\mathbb{1}(\cdot)$ is the indicator function. For background tokens, we suppress their updates through a gradient blocking operation:

$$X'_{b,i} = \begin{cases} X_{b,i}, & M_{b,i} = 0 \\ \text{detach}(X_{b,i}), & M_{b,i} = 1 \end{cases} \tag{6}$$

where detach$(\cdot)$ is the gradient truncation operator. Subsequently, we add learnable positional encodings $P_z \in \mathbb{R}^{N_z \times D}$ and $P_x \in \mathbb{R}^{N_x \times D}$ to the template features $Z \in \mathbb{R}^{B \times N_z \times D}$ and the updated search features X', respectively, to preserve spatial structure information. Finally, the classification token $c \in \mathbb{R}^{B \times 1 \times D}$ is concatenated with both to form the complete input sequence:

$$\tilde{X} = [c; Z + P_z; X' + P_x] \in \mathbb{R}^{B \times (1+N_z+N_x) \times D} \tag{7}$$

This method actively attenuates background interference during the feature extraction stage, enabling subsequent attention computation to focus more on target regions while reducing unnecessary parameter updates through gradient blocking.

3.2.3 Language-Guided Cross-Modal Feature Fusion

In multi-modal tracking, visual-language feature fusion efficiency directly impacts performance. Literature[27] notes that existing trackers use complex fusion modules with parameter redundancy and high computational overhead, failing UAV real-time requirements. Their dynamic sparse learning mechanism proves that selectively retaining important tokens while suppressing irrelevant information effectively improves fusion efficiency.

Based on the above observation, we design a hierarchical dynamic fusion mechanism, which consists of M deep encoder layers. For language features $T \in \mathbb{R}^{B \times L_t \times D}$, due to their relatively small sequence size, directly pruning tokens yields limited benefits. However, dimensional redundancy exists in language features, so semantic screening is performed at the channel dimension. In the first layer, we compute the interactive response between the mean of visual features and language features:

$$C = \left(\frac{1}{L_v} \sum_{i=1}^{L_v} V_{i,:} \right) \odot T \in \mathbb{R}^{B \times L_t \times D} \tag{8}$$

where $V \in \mathbb{R}^{B \times L_v \times D}$ represents the visual features, $L_v = 1 + N_z + N_x$, N_z is the number of template tokens, and N_x is the number of search tokens. This response measures the correlation between each language channel and the visual features. We retain the top $k = 3/4$ channels with the strongest responses, construct a channel mask M_{chan}, and apply constraints to the language features in subsequent layers:

$$T' = T \odot M_{\text{chan}} \tag{9}$$

For the search region $V_x \in \mathbb{R}^{B \times N_x \times D}$, we propose a progressive token masking strategy based on attention responses. We extract the attention response $a_L \in \mathbb{R}^{B \times N_x}$ of template tokens to the search region, set the masking ratio $\alpha_L = \max(0, L \times 0.1)$, retain the top $k_L = \lfloor N_x(1 - \alpha_L) \rfloor$ tokens with the strongest responses, construct a token mask M_{token}, and apply constraints to the search region features:

$$V_x' = V_x \odot M_{\text{token}} \tag{10}$$

Additionally, in each layer of vision-language feature fusion, modality embeddings $m_v, m_t \in \mathbb{R}^D$ are added to the visual features V and the filtered language features T', respectively, and then concatenated:

$$E = [V + m_v; T' + m_t] \in \mathbb{R}^{B\times(L_v+L_t)\times D} \tag{11}$$

Through visually-guided semantic screening of language features in the channel dimension and attention-guided spatial focusing of visual features in the token dimension, this deep fusion significantly enhances the discriminative capability and computational efficiency of the multi-modal tracker.

3.2.4 Target Feature Decoding and Localization.

We achieve target localization through modality-adaptive bounding box prediction. In the localization process, visual features $V \in \mathbb{R}^{B\times L_v\times D}$ and language features $T \in \mathbb{R}^{B\times L_t\times D}$ are first concatenated into a multi-modal token group. Based on the flag vectors $f \in \mathbb{R}^B$ output from the same layer in the network, corresponding to target, distractor, and background respectively, the corresponding modality token $p \in \mathbb{R}^{B\times D}$ is dynamically selected from the token group, enabling adaptive selection of the most suitable feature representation according to the current target state.

Subsequently, a distribution-aware cross-attention module generates prompt features $H \in \mathbb{R}^{B\times 2\times D}$, partitioning the context into target and background categories. The similarity between search features $X \in \mathbb{R}^{B\times N_x\times D}$ (where N_x is the number of search tokens) and prompt features is computed, yielding contrastive scores $c \in \mathbb{R}^{B\times N_x}$ that measure the matching degree of each search position with the target.

Building upon this, the selected modality token p is element-wise multiplied with the search features, and convolutional layers are used to predict the classification score map $C \in \mathbb{R}^{B\times H\times W}$, offsets $O \in \mathbb{R}^{B\times 2\times H\times W}$, and bounding box sizes $B \in \mathbb{R}^{B\times 2\times H\times W}$. Combined with the predefined coordinate grid $P \in \mathbb{R}^{B\times 2\times N}$, candidate boxes for each position are computed as:

$$Q = \left[\frac{P + O}{S}; B\right] \in \mathbb{R}^{B\times N\times 4} \tag{12}$$

where S is the stride of the feature map relative to the input image, and the four dimensions of Q correspond to the predicted box's center coordinates, width, and height. Finally, the candidate box with the highest joint weighting of classification score and contrastive score is selected as the final prediction.

4 Experiments

4.1 Experimental Setup

Dataset. To ensure objective comparability with existing natural language-based tracking methods, we employ seven public datasets for joint training, including OTB99[23], TNL2K[22], RefCOCOg[25], GOT10K[37], and LaSOT[24]. For evaluation, we select three datasets: OTB99, LaSOT, and TNL2K, randomly sampling 6,400 points per epoch for performance testing.

Implementation Details. All experiments are conducted on an RTX 3090. We employ the Adam optimizer for parameter optimization with an initial learning rate of 0.0004, dynamically adjusted using a cosine annealing strategy. Input search images are

Table 1 Comparison of tracking performance under different reference modes with natural language. The best and second-best results are marked in bold and underlined, respectively.

Method	TNL2K		LaSOT		OTB99	
	AUC	**PRE**	**AUC**	**PRE**	**AUC**	**PRE**
Method A	0.85	0.90	0.80	0.85	0.75	0.80
Method B	0.88	0.92	0.82	0.87	0.78	0.82

randomly cropped into 256 patches, while template images are divided into 64 patches. The maximum length of language input is set to 40, with encoder layers configured as N=6 and M=6. Additionally, our VQATrack is implemented on an AirSim-based UAV equipped with a forward-facing RGB vision sensor. The camera resolution is 1920×1080, installed at a 10-degree downward pitch, positioned 0.3 meters forward and 0.1 meters below the UAV's center of gravity. These detailed specifications facilitate quick reproduction and sensor parameter adjustment by researchers for different research needs.

4.2 Comparative Analysis of Language Generation Results

To meet the initialization requirements of UAV vision-language tracking tasks and verify the impact of different input references on VQATrack's target tracking accuracy, we compare three distinct input modes using Qwen3-VL-4B-Instruct-FP8: (1) target template only, (2) search image with bounding box, and (3) simultaneous input of search image and target template.

Subsequently, we conducted tests on three datasets (TNL2K, LaSOT, and OTB99) using the same prompt, with experimental results shown in Table 1. It can be observed that using only the target template or the search image with bounding box causes the model to hallucinate, affecting tracking accuracy. In contrast, using both the search image and target template effectively replaces traditional manual descriptions, improving the efficiency of language description generation while ensuring semantic accuracy. Furthermore, our findings are largely consistent with those reported in literature[38], verifying that the visual question answering model has minimal impact on the accuracy of language description generation.

4.3 State-of-the-Art Comparisons

To validate the performance of our proposed VQATrack, we compare our method across multiple dimensions. Table 2 presents quantitative comparison results between VQATrack and existing mainstream trackers. On the TNL2K dataset, VQATrack achieves 66.2 AUC and 69.7 PRE under the NL+BB input mode, outperforming recent methods such as UVLTrack-B. On the LaSOT dataset, VQATrack also achieves the best performance, with AUC and PRE reaching 72.3 and 80.0 respectively, showing improvements over the second-best methods. On the OTB99 dataset, VQATrack attains 69.5 precision, performing on par with mainstream methods.

Subsequently, Figure 3 analyzes the tracking performance of different methods under various challenging attributes. VQATrack achieves the best scores on most challenging attributes, particularly in complex scenarios such as target deformation, background clutter, and illumination variation.

Table 2 AUC and PRE of different methods on TNL2K, LaSOT and OTB99 datasets. BB and NL denote bounding box and natural language.

Method	Initialize	TNL2K		LaSOT		OTB99	
		AUC	PRE	AUC	PRE	AUC	PRE
SiamRPN++ [39]	BB	41.3	41.2	49.6	49.1	-	-
Ocean [6]	BB	38.4	37.7	56.0	56.6	-	-
SiamFC++ [40]	BB	38.6	36.9	54.4	54.7	-	-
TrDiMP [41]	BB	52.3	52.8	63.9	61.4	-	-
TransT [42]	BB	50.7	51.7	64.9	69.0	-	-
SwinTrack-B [43]	BB	54.8	53.8	69.3	-	-	-
UVLTrack-B [12]	BB	62.7	65.4	69.4	74.9	66.9	**81.7**
SUTrack-B224 [44]	BB	65.0	67.9	73.2	**80.5**	-	-
VQATrack (ours)	BB	**65.2**	**68.9**	**74.2**	79.2	**70.0**	80.6
TNLS-III [8]	NL+BB	-	-	-	-	55.0	72.0
RVTNLN [45]	NL+BB	25.0	27.0	50.0	56.0	67.0	73.0
RTTNLD [46]	NL+BB	25.0	27.0	35.0	35.0	61.0	79.0
TNL2K-2 [22]	NL+BB	41.7	42.0	51.0	55.0	68.0	88.0
JointNLT [47]	NL+BB	56.6	57.9	58.8	62.3	63.6	87.1
UVLTrack-B [12]	NL+BB	63.1	66.7	69.4	75.9	69.3	**89.9**
VQATrack (ours)	NL+BB	**66.2**	**69.7**	**72.3**	**78.6**	**69.5**	89.5

To qualitatively evaluate VQATrack's performance, we present visualization results under various UAV scenes in Figure 5. Compared with existing methods, VQATrack demonstrates superior localization accuracy and adapts well to appearance variations, maintaining stable tracking even under occlusion and similar object interference.

From the perspective of temporal tracking stability, Figure 6 compares the IoU score variations of VQATrack and existing methods across complete sequences. VQATrack maintains higher IoU levels throughout the entire tracking process, exhibits smaller performance degradation, and demonstrates faster recovery speed, reflecting stronger temporal robustness.

Finally, we conduct efficiency analysis in Table 3, comparing FLOPs and inference time across models. While separated methods like VLTVG+STARK suffer from efficiency bottlenecks with nearly 170M parameters, our unified VQATrack achieves optimal performance with only 108.4M parameters—35% fewer than UVLTrack. GPU memory consumption is reduced to 65%, and inference speed reaches 53.2 FPS,

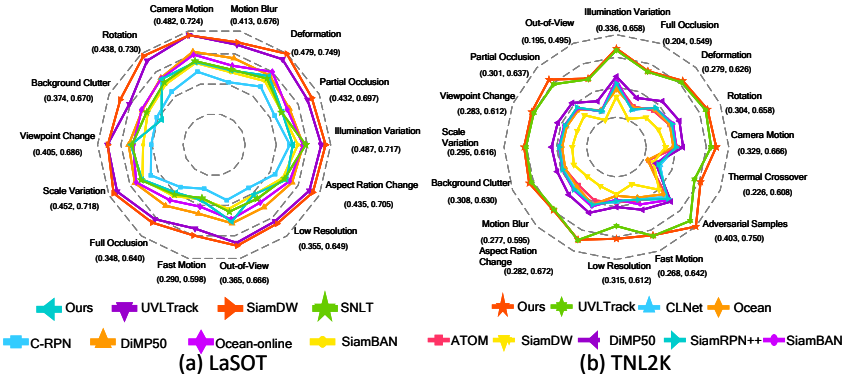


Fig. 4 Comparison of VQATrack with other state-of-the-art trackers. AUC scores for each challenging attribute are evaluated on the TNL2K and LaSOT datasets under natural language mode.

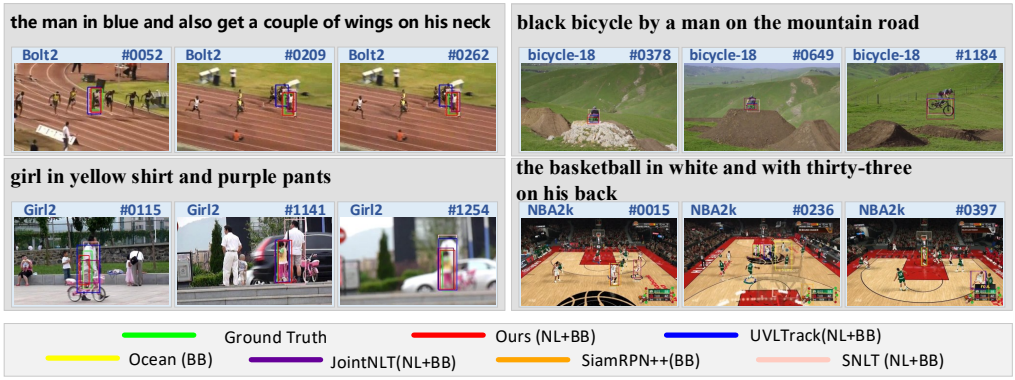
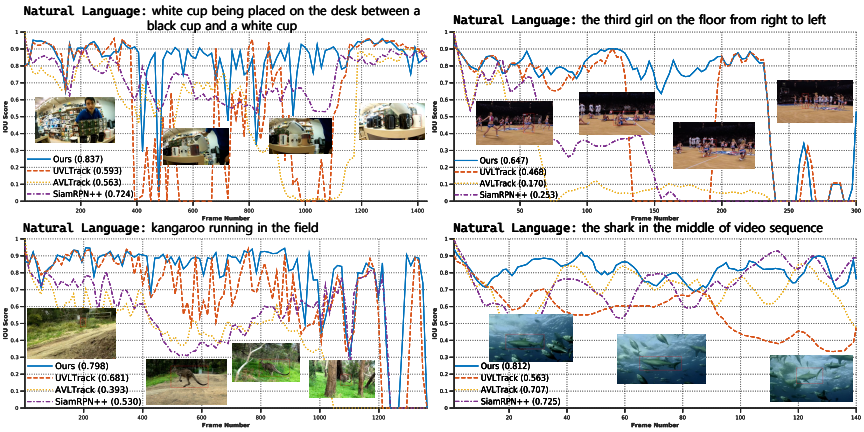


Fig. 5 Qualitative visualization comparison of VQATrack with other advanced models.

demonstrating an excellent balance between lightweight design and state-of-the-art accuracy.

4.4 Comparison of ablation experiments

To validate the effectiveness of each module, we conducted ablation experiments on the LaSOT and TNL2K datasets. After introducing the Efficient Encoder to the baseline model, inference speed increased to 55 FPS, memory usage decreased to 75%, and the average AUC under NL+BB mode improved by 1.5%. With the addition of visual token masking, FPS further increased to 58, and memory usage dropped to 70%. Finally, after incorporating feature interaction fusion, the model achieved optimal AUC performance under natural language mode, with an average improvement of 3.1% over the baseline. The ablation results validate the progressive contribution of each module to both efficiency and accuracy.



(a) LASOT (b) TNL2K

Fig. 6 Comparison of IoU score trends between VQATrack and different models under NL+BB mode.

Table 3 Comparison of Separate and Joint Methods. FLOP and inference times for Tracking in NL+BB mode.

		Separated Model			Joint Model		
		VLTVG[48] +STARK[49]	VLTVG[48] +OSTrack[50]	SepRM [47]	JointNLT	UVLTrack	VQATrack (ours)
FLOPs	Tracking	20.4G	48.3G	38.5G	42.0G	43.7G	**53.2G**
FPS	Tracking	43.6	120.4	48.5	39.2	44.1	**62.0**
AUC	LaSOT	44.6	52.4	51.8	58.8	69.4	**72.3**
	TNL2K	37.3	39.9	41.9	56.6	63.1	**66.2**
Params	Total	169.8M	214.7M	214.4M	153.0M	167.7M	**108.4M**
GPU Memory	Total	-	-	-	88%	90%	**65%**

4.5 Autonomous UAV Target Tracking and Control

To address the requirements of target depth perception and real-time obstacle avoidance in autonomous UAV tracking tasks, we propose a vision-based autonomous UAV tracking and control method. First, the tracking model outputs the target bounding box in the image plane, providing spatial reference for subsequent depth sampling. However, traditional RGB cameras lack depth information, while equipping LiDAR or depth cameras increases payload and cost. To address this, we employ the Depth Anything V2 (ViT-L)[51] depth estimation method. The model estimates pixel-level depth from RGB images and suppresses inter-frame jitter in the estimation results through an exponential smoothing mechanism. Subsequently, for UAV control, we design velocity commands tailored for UAV target tracking tasks to ensure that the UAV maintains a stable tracking distance from the target. Additionally, as shown in Figure 7, we adopt a simple and reliable obstacle avoidance strategy to reduce the

Table 4 Ablation study of VQATrack with different components.

Method	FPS	GPU Memory	LaSOT				TNL2K			
			BB		NL+BB		BB		NL+BB	
			AUC	PRE	AUC	PRE	AUC	PRE	AUC	PRE
Baseline	44	90%	69.4	74.9	69.4	75.9	62.7	65.4	63.1	66.7
+ Language Self-Generation	44	90%	63.3	66.9	69.4	75.9	69.4	89.9	63.3	66.9
+ Efficient Encoder	55	75%	72.0	77.2	70.6	77.7	63.9	67.3	64.8	68.4
+ Visual Token Masking	58	70%	73.4	78.1	71.0	78.1	64.6	68.0	65.4	69.0
+ Feature Interactive Fusion	**62**	**65%**	**74.2**	**79.2**	**72.3**	**78.6**	**65.2**	**68.9**	**66.2**	**69.7**

real-time computational burden during UAV flight. The central region of the depth map is divided into three sub-regions: left, center, and right. The minimum depth value for each region is calculated, and the lateral offset direction is determined based on depth comparison between the left and right regions.

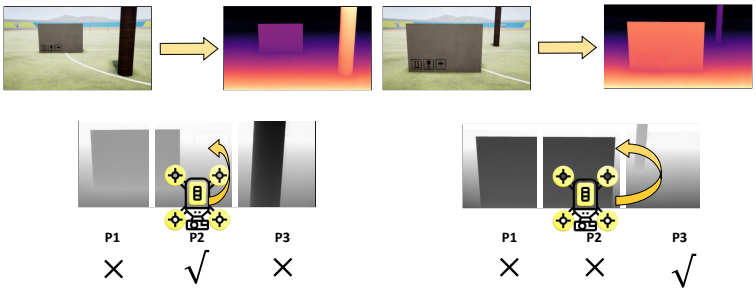


Fig. 7 UAV obstacle avoidance strategy

UAV Simulation Results. We validated our UAV tracking strategy in simulation. As shown in Figure 8(a), VQATrack accurately outputs target descriptions, maintains smooth trajectories, and achieves stable obstacle avoidance.

In Figure 8(b), under challenging conditions including large target movements, full occlusion, and interfering figures, the UAV maintains stability without target loss. It accurately maintains preset distance and keeps the target centered, demonstrating advantages in real-time performance and stability.Finally, tests show that our method can run stably at 10 FPS on the NVIDIA Jetson Orin NX embedded platform, meeting the real-time requirements.

5 Conclusion

This paper proposes VQATrack, a lightweight vision-language tracking and control method for UAV embedded platforms, and builds a full-link closed-loop simulation

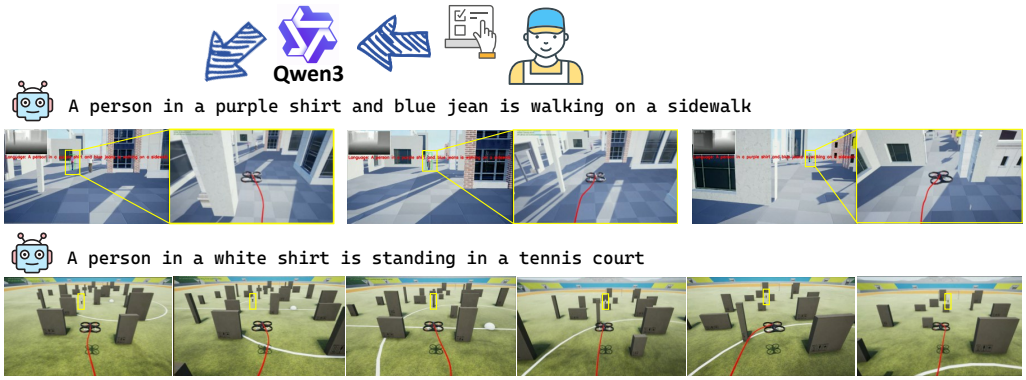


Fig. 8 UAV autonomous tracking and obstacle avoidance simulation results

system from perception to control. The method eliminates manual annotation via a visual question answering-based language generation module, achieves lightweight tracking through hierarchical gradient freezing, token masking, and dynamic fusion, and integrates depth estimation with obstacle avoidance. Experiments show VQATrack runs at 10 FPS on NVIDIA Jetson Orin NX, yielding a speed gain of 8 FPS with improved accuracy. It enables stable tracking and autonomous obstacle avoidance in dense obstacle scenarios. Future work will focus on model compression, aerial depth estimation, and end-to-end dynamic obstacle avoidance.

6 Discussion

While VQATrack demonstrates promising performance, there are still some limitations that warrant discussion. First, the VQA-based language generation module relies on a pre-trained model, which might sometimes generate descriptions that are not perfectly aligned with the tracking target in highly cluttered environments. Second, the depth estimation module assumes a relatively flat ground plane; thus, its accuracy may degrade in complex terrains with significant elevation changes. Finally, although our closed-loop simulation validates the system's effectiveness, there remains a domain gap between the simulated environment and real-world flight dynamics, such as unpredictable wind disturbances and physical sensor noise. In future work, we plan to deploy the system on a real UAV to further test its robustness in real-world scenarios.

Supplementary information. Not applicable.

Acknowledgements. This work was supported by (1) The Shaanxi Natural Science Basic Research Project [Grant numbers 2024JC-YBQN-0660]. (2) The Key Laboratory of Marine Environmental Survey Technology and Application, Ministry of Natural Resources [Grant numbers MESTA-2024-B002]. (3) The Key Laboratory of Intelligent Optoelectronics of the Ministry of Education [Grant numbers ZNGD20231010].

Declarations

Funding. This work was supported by the Shaanxi Natural Science Basic Research Project [Grant numbers 2024JC-YBQN-0660], the Key Laboratory of Marine Environmental Survey Technology and Application, Ministry of Natural Resources [Grant numbers MESTA-2024-B002], and the Key Laboratory of Intelligent Optoelectronics of the Ministry of Education [Grant numbers ZNGD20231010].

Conflict of interest. The authors declare that they have no competing interests or conflict of interest.

Ethics approval and consent to participate. Not applicable. This study did not involve human participants or animal experiments.

Consent for publication. Not applicable. This manuscript does not contain any individual person's data in any form.

Data availability. The datasets generated during and/or analyzed during the current study are available in the public repositories. Specifically, the LaSOT, TNL2K, and OTB99 datasets are publicly available and were used for evaluation in this paper.

Materials availability. Not applicable.

Code availability. The source code is publicly available at: https://github.com/yll1214/VQATrack

Author contribution. Lingling Yang contributed to the conceptualization, methodology, and writing of the original draft. Liqiang Liu and Yanfang Fu contributed to the software implementation, validation, and data curation. Anyuan Xie and Yujia Li contributed to the formal analysis and visualization. Liqiang Liu and Ningchao Zhang supervised the project, acquired the funding, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

References

[1] Wen, J., Hu, Y., Qiu, S., Huang, J., Chu, X.: Sdg-track: A heterogeneous observer-follower framework for high-resolution uav tracking on embedded platforms. ArXiv **abs/2512.04883** (2025)

[2] Fu, C., Lei, X., Zuo, H., Yao, L., Zheng, G.-Z., Pan, J.: Progressive representation learning for real-time uav tracking. 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 5072–5079 (2024)

[3] Müller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018, pp. 310–327. Springer

[4] Hong, L., Yan, S., Zhang, R., Li, W., Zhou, X., Guo, P., Jiang, K., Chen, Y., Li, J., Chen, Z., Zhang, W.: Onetracker: Unifying visual object tracking with foundation

models and efficient tuning. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 19079–19091 (2024)

[5] Danelljan, M., Gool, L.V., Timofte, R.: Probabilistic regression for visual tracking. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7181–7190 (2020)

[6] Zhang, Z., Peng, H.: Ocean: Object-aware anchor-free tracking. ArXiv **abs/2006.10721** (2020)

[7] Wang, Y., Hu, S., Jia, S., Xu, P., Ma, H., Ma, Y., Zhang, J., Lu, X., Zhao, X.: Soi is the root of all evil: Quantifying and breaking similar object interference in single object tracking. ArXiv **abs/2508.09524** (2025)

[8] Li, Z., Tao, R., Gavves, E., Snoek, C.G.M., Smeulders, A.W.M.: Tracking by natural language specification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6495–6503

[9] Feng, X., Hu, S., Li, X., Zhang, D., Wu, M., Zhang, J., Chen, X., Huang, K.: Atctrack: Aligning target-context cues with dynamic target states for robust vision-language tracking. https://api.semanticscholar.org/CorpusID:280323608

[10] Maurya, A., Ye, J., Rafique, M.M., Cappello, F., Nicolae, B.: Breaking the memory wall: A study of i/o patterns and gpu memory utilization for hybrid cpu-gpu offloaded optimizers. In: Proceedings of the 14th Workshop on AI and Scientific Computing at Scale Using Flexible Computing Infrastructures, pp. 9–16

[11] Tan, M., Le, Q.V.: Efficientnetv2: Smaller models and faster training. (2021). https://api.semanticscholar.org/CorpusID:232478903

[12] Ma, Y., Tang, Y., Yang, W., Zhang, T., Zhang, J., Kang, M.: Unifying visual and vision-language tracking via contrastive learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 4107–4116

[13] Cao, Z., Fu, C., Ye, J., Li, B., Li, Y.: Siamapn++: Siamese attentional aggregation network for real-time uav tracking. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 3086–3092 (2021)

[14] Yao, L., Fu, C., Li, S., Zheng, G.-Z., Ye, J.: Sgdvit: Saliency-guided dynamic vision transformer for uav tracking. 2023 IEEE International Conference on Robotics and Automation (ICRA), 3353–3359 (2023)

[15] Wu, Y., Li, Y., Liu, M., Wang, X., Yang, X., Ye, H., Zeng, D., Zhao, Q., Li, S.: Learning an adaptive and view-invariant vision transformer for real-time uav tracking. IEEE Transactions on Circuits and Systems for Video Technology **36**, 2403–2418 (2024)

[16] Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016, pp. 445–461. Springer

[17] Li, B., Fu, C., Ding, F., Ye, J., Lin, F.: All-day object tracking for unmanned aerial vehicle. IEEE Transactions on Mobile Computing **22**, 4515–4529 (2021)

[18] Yao, L., Fu, C., Wang, Y., Zuo, H., Lu, K.: Enhancing nighttime uav tracking with light distribution suppression. 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 5902–5909 (2024)

[19] Zhao, X., Hu, S., Wang, Y., Zhang, J., Hu, Y., Liu, R., Ling, H., Li, Y., Li, R., Liu, K., Li, J.: Biodrone: A bionic drone-based single object tracking benchmark for robust vision. International Journal of Computer Vision **132**(5), 1659–1684 (2024) https://doi.org/10.1007/s11263-023-01937-0

[20] Wu, Y., Wang, X., Yang, X., Liu, M., Zeng, D., Ye, H., Li, S.: Learning occlusion-robust vision transformers for real-time uav tracking. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 17103–17113 (2025)

[21] Chen, C., Liang, S., Guan, R., Sun, X., Zhao, H., Jiang, H., Huang, T., Ding, H., Han, Q.-L.: Aerialmind: Towards referring multi-object tracking in uav scenarios. ArXiv **abs/2511.21053** (2025)

[22] Wang, X., Shu, X., Zhang, Z., Jiang, B., Wang, Y., Tian, Y., Wu, F.: 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

[23] Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(9), 1834–1848 (2015) https://doi.org/10.1109/TPAMI.2014.2388226

[24] Fan, H., Bai, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Harshit, Huang, M., Liu, J., *et al.*: Lasot: A high-quality large-scale single object tracking benchmark. International Journal of Computer Vision **129**, 439–461 (2021)

[25] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11–20

[26] Zhang, C., Huang, G., Liu, L., Huang, S., Yang, Y., Wan, X., Ge, S., Tao, D.: 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) https://doi.org/10.1109/TPAMI.2022.3232854

[27] Xue, Y., Zhong, B., Jin, G., Shen, T., Tan, L., Li, N., Zheng, Y.: Avltrack:

Dynamic sparse learning for aerial vision-language tracking. IEEE Transactions on Circuits and Systems for Video Technology **35**(8), 7554–7567 (2025) https://doi.org/10.1109/TCSVT.2025.3549953

[28] Zhang, Y., Rui, X., Song, W.: A uav-based multi-scenario rgb-thermal dataset and fusion model for enhanced forest fire detection. Remote Sensing **17**(15), 2593 (2025)

[29] Xue, Y., Zhang, J., Lin, Z., Li, C., Huo, B., Zhang, Y.: Siamcaf: Complementary attention fusion-based siamese network for rgbt tracking. Remote Sensing **15**(13), 3252 (2023)

[30] Gao, Z., Li, D., Wen, G., Kuai, Y., Chen, R.: Drone based rgbt tracking with dual-feature aggregation network. Drones **7**(9), 585 (2023)

[31] Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X.-H., Cheng, Z., Deng, L., Ding, W., Fang, R., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, Q., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L.Y., Ren, X., Ren, X.-y., Song, S., Sun, Y.-c., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. ArXiv **abs/2511.21631** (2025)

[32] Wang, Z., Lin, Z., Yang, Y., Fu, H., Ye, D.: Seenav-agent: Enhancing vision-language navigation with visual prompt and step-level policy optimization. ArXiv **abs/2512.02631** (2025)

[33] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and Short Papers), pp. 4171–4186

[34] He, K., Chen, X., Xie, S., Li, Y., Dollr, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009. á

[35] Deng, A., Cao, T., Chen, Z., Hooi, B.: Words or vision: Do vision-language models have blind faith in text? 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3867–3876 (2025)

[36] Cai, M., Yang, J., Gao, J., Lee, Y.J.: Matryoshka multimodal models. ArXiv **abs/2405.17430** (2024)

[37] Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for

generic object tracking in the wild. IEEE transactions on pattern analysis and machine intelligence **43**(5), 1562–1577 (2019)

[38] Li, X., Feng, X., Hu, S., Wu, M., Zhang, D., Zhang, J., Huang, K.: Dtllm-vlt: Diverse text generation for visual language tracking based on llm. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 7283–7292 (2024)

[39] Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4282–4291

[40] Xu, Y., Wang, Z., Li, Z., Yuan, Y., Yu, G.: Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12549–12556

[41] Wang, N., Zhou, W., Wang, J., Li, H.: 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

[42] Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8126–8135

[43] Lin, L., Fan, H., Zhang, Z., Xu, Y., Ling, H.: Swintrack: A simple and strong baseline for transformer tracking. Advances in Neural Information Processing Systems **35**, 16743–16754 (2022)

[44] Chen, X., Kang, B., Geng, W., Zhu, J., Liu, Y., Wang, D., Lu, H.: Sutrack: Towards simple and unified single object tracking. ArXiv **abs/2412.19138** (2024)

[45] Feng, Q., Ablavsky, V., Bai, Q., Sclaroff, S.: Robust visual object tracking with natural language region proposal network. arXiv preprint arXiv:1912.02048 **1**(7), 8 (2019)

[46] Feng, Q., Ablavsky, V., Bai, Q., Li, G., Sclaroff, S.: Real-time visual object tracking with natural language description. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 700–709

[47] Zhou, L., Zhou, Z., Mao, K., He, Z.: 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

[48] Yang, L., Xu, Y., Yuan, C., Liu, W., Li, B., Hu, W.: Improving visual grounding with visual-linguistic verification and iterative reasoning. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9499–9508

[49] Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10448–10457

[50] Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: European Conference on Computer Vision, pp. 341–357

[51] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. ArXiv **abs/2406.09414** (2024)