Context-Aware Integration of Language and Visual References for Natural Language Tracking

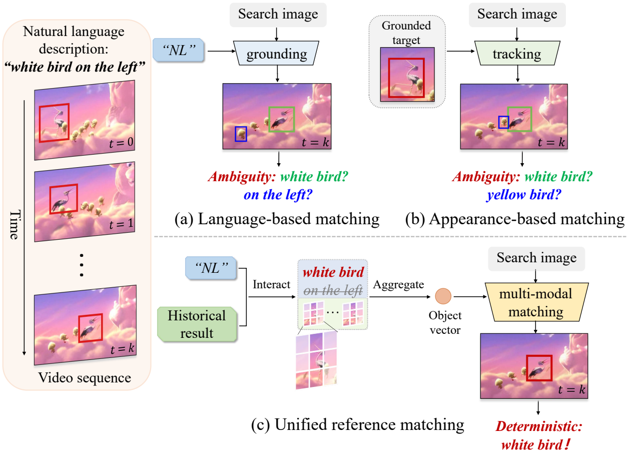
Yanyan Shao[1] Shuting He[2] Qi Ye [3*] Yuchao Feng[1] Wenhan Luo[4] Jiming Chen[1,3]
[1] Zhejiang University of Technology [2] Nanyang Technological University [3] Zhejiang University
[4] The Hong Kong University of Science and Technology
{shaoyanyan,fyc}@zjut.edu.cn, shuting.he@ntu.edu.sg, {qi.ye,cjm}@zju.edu.cn, whluo@ust.hk

Abstract

Tracking by natural language specification (TNL) aims to consistently localize a target in a video sequence given a linguistic description in the initial frame. Existing methodologies perform language-based and template-based matching for target reasoning separately and merge the matching results from two sources, which suffer from tracking drift when language and visual templates miss-align with the dynamic target state and ambiguity in the later merging stage. To tackle the issues, we propose a joint multi-modal tracking framework with 1) a prompt modulation module to leverage the complementarity between temporal visual templates and language expressions, enabling precise and context-aware appearance and linguistic cues, and 2) a unified target decoding module to integrate the multi-modal reference cues and executes the integrated queries on the search image to predict the target location in an end-to-end manner directly. This design ensures spatio-temporal consistency by leveraging historical visual information and introduces an integrated solution, generating predictions in a single step. Extensive experiments conducted on TNL2K, OTB-Lang, LaSOT, and RefCOCOg validate the efficacy of our proposed approach. The results demonstrate competitive performance against state-of-the-art methods for both tracking and grounding. Code is available at https://github.com/twotwo2/QueryNLT

1. Introduction

Tracking by natural language specification (TNL) aims to localize the target object in a video sequence based on a given language description on the initial frame. It offers a more user-friendly interaction to specify the target object compared to traditional tracking-by-bounding-box methods [2, 9, 15, 37], which has a wide range of applications in video surveillance, robotics, and autonomous vehicles.



(a) Language-based matching (b) Appearance-based matching

(c) Unified reference matching

Figure 1. Given a video sequence, the tracking object is characterized as *"white bird on the left"* of the initial frame. Existing two-step approaches separately perform language-search matching (a) and appearance-search matching (b). However, *"on the left"* which is inconsistent with the current target and the background contained in the grounded target may confuse the identification of the target. In contrast, our QueryNLT (c) forms a dynamic and context-aware query for target localization by integrating visual and language references. (Zoom in for a better view).

Previous research efforts [5, 6, 18, 30, 35] generally divide language-guided tracking into two fundamental subtasks: visual grounding and visual tracking. These studies initially localize the target object solely based on the given language description, i.e. visual grounding, and the grounded target serves as the visual template to establish correspondence with the search image, i.e. visual tracking. The final results are derived through the amalgamation of the outcomes obtained from both visual grounding and visual tracking. Despite the significant success, these approaches typically process the language and template independently until merging their matching results, which may lead to ambiguity in target identification. As illustrated in Fig. 1(a), due to the target's movement, the initially pro-

[*] Corresponding author

vided language description (*"white bird on the left"*) may no longer align with the current state of the target (the bird moves to the middle at k^{th} frame). This misalignment confuses the tracker's judgment of whether to focus on *"white bird"* or *"on the left"* during the matching process. What is more, occlusions of objects may bring background clutters into the template. As shown in Fig. 1(b), a small yellow bird is contained in the template, which may further interfere with the identification of the target "*white bird*". The late fusion at the resultant level makes it difficult for the tracker to discern which candidate object is the real target, thus leading to tracking drift.

Based on the observation, we argue that the language description and the visual template are complementary and combining these two for matching contributes to a comprehensive understanding and perception of the target. To form the accuracy and context-aware target information as guidance, we propose a multi-modal prompt modulation module to filter out descriptions in the initial verbal reference and the visual reference accumulated in the tracking history that does not align with the current state. As illustrated in Fig. 1(c), the historical results embed the target's motion information, which helps filter out status descriptions that fail to align with the actual target in the language expression. For instance, *"on the left"* referring to a small yellow bird rather than the true target, should be removed. Simultaneously, the categorization of the target, as depicted in the language description, serves as a reliable cue for filtering out extraneous background features in the visual template. Specifically, patches belonging to *"white bird"* are assigned high attention weights, and patches belonging to the yellow bird and background are masked. The revised language description collaborates with the accurate visual template to help point to the true target in the challenging scene.

Afterward, we present a query-based target decoding module that jointly establishes the correspondence between the multi-modal references with the search image in a one-step fashion. The key insight is to consider the language-based matching subtask and the appearance-based matching subtask as a unified instance-level retrieval problem. To achieve this, this module comprises a multi-modal query generator that aggregates visual and verbal cues into a holistic object vector, and a query-based target locator that establishes the correspondence between the query vector and the search image for target retrieval. Compared with the previous works that need post-processing for merging results, it can directly predict the target location in an end-to-end manner. The prompt modulation module along with the target decoder module forms a unified framework to utilize the verbal and visual reference for natural language tracking. With such a design, our proposed framework not only effectively improves the target discrimination through integrated perception, but also ensures the spatio-temporal consistency by forming context-aware query information.

We validate the effectiveness of our proposed framework through comprehensive evaluations on three tracking benchmarks and a grounding benchmark, including TNL2K [30], OTB-Lang [18, 31], LaSOT [4], and RefCOCOg [23]. Without bells and whistles, our QuertNLT achieves competitive performance compared with state-of-the-art trackers. Our main contributions are as follows.

- We propose a novel framework for the natural language tracking task, termed QueryNLT. This framework integrates diverse modal references for target modeling and matching, fostering a holistic understanding of the target and improving discrimination capabilities.
- We propose a prompt modulation module that explores the complementarity of multi-modal reference to eliminate the inconsistent descriptions in the reference, generating precise and context-aware cues for target retrieval.
- We conduct comprehensive experiments on three challenging natural language tracking datasets and a visual grounding dataset, validating the efficacy of our proposed framework. The results showcase its robust performance and suitability for diverse tracking scenarios.

2. Related Work

In this work, we aim to improve the performance of language-guided tracking by joining heterogeneous visual and language references. In the following, we will discuss related work that explores the utilization of these two heterogeneous references in existing language-guided tracking approaches, as well as how language-assistant target tracking approaches to underscore the potential benefits of the multi-modal tracking approach.

2.1. Language-guidance Object Tracking

The emerging field of tracking by natural language specification (TNL) has garnered significant attention in recent years. It presents a unique approach to precisely localizing target objects within video sequences based on corresponding language descriptions. As the pioneering work in this area, Li *et al.* [18] first define the task of tracking by natural language specification and demonstrate the feasibility of language description replacing bounding boxes to specify targets. Subsequently, Yang *et al.* [35] and Feng *et al.* [6] share the same solution that divides this task into two subtasks: a grounding task solely relying on language to find the target and a tracking task based on the grounding results as the template. To better utilize the semantic information of the target during the tracking phase, [35] simultaneously performs visual matching based on the history of grounded objects, as well as performs grounding based on the language query for each subsequent frame. Besides, they propose an integration module to combine the prediction results of both processes adaptively. With the help of the region proposal network, [6] follows the tracking-by-detection formulation, leveraging language to select the most suitable proposal as the target template for tracking.

In order to accelerate research for TNL, Wang *et al.* [30] release a new benchmark and propose an adaptive switch framework that performs global search with language reference or local matching with visual template reference. While all of these approaches have made great progress, however, the grounding module used to initialize the template and the subsequent tracking used for tracking are separate, and cannot be trained end to end. Recently, Zhou *et al.* [41] introduce a joint framework to replace a separate framework aiming at linking the language and template reference. However, it overlooks that the language expression may be inconsistent with the current tracking scene, which may cause references to be ambiguous. In this paper, we present a novel and effective framework that takes into account both linguistic descriptions and visual template information to improve target discrimination, while utilizing the complementary nature of heterogeneous information to form more accurate target reference information.

2.2. Language-assisted Object Tracking

Different from the language-guidance tracking approaches the target object is specified only by the language description of the first frame, the tracking object of the language-assisted approaches [7, 8, 10, 16, 27, 29, 40] is specified by both box and language. With language description as an auxiliary cue, these works often focus on transforming a traditional box-guided tracking approach into a multi-modal target tracking approach.

Some work has been done to improve the performance of the tracker, in terms of improving visual feature representation [10, 29] and enhancing the matching associations with the search image [7, 16]. On the one hand, Feng *et al.* [7] propose to perform symmetrical language-based matching alongside template-based matching [14, 15], where the results of both branches are weighted to obtain the final result. On the other hand, Guo *et al.* [10] treat language as a selector to reweight visual features and enhance visual feature representation through neural architecture search technology [25, 43]. In contrast to the aforementioned methods where the language description is provided by the user, Li *et al.* [16] propose to automatically generate the corresponding semantic descriptions based on the input template. Taking advantage of the text-image alignment capability of CLIP models [24], [16] designs to select the corresponding semantic descriptions from predefined attributes can be used as complementary descriptions. These methods demonstrate that verbal cues alongside visual cues significantly enhance the overall understanding of the target, thus improving target discrimination.

3. Method

3.1. Overview

Our goal is to consistently and accurately localize the target within a video sequence, which is specified by language description. The main observation of our work is that the dynamic visual cue and the language expression provide complementary information that enhances target perception and discrimination. Diverging from previous methods [18, 29, 30] that employ separate networks for language-based and template-based matching, our proposed QueryNLT treats these two sub-tasks as an instance retrieval problem. To this end, we propose a unified multi-modal matching network for language-guided tracking.

The framework of QueryNLT is depicted in Fig. 2. During the tracking phase, we collect the appearance feature h_a and positional feature h_p of the target based on the historical localization results of the target and store in a template memory $\mathcal{M} = \{h_a, h_p\}$ as the dynamic visual reference. Given a search image I_s and a language description $\mathcal{D}$, we first employ a feature extraction module (in section 3.2) to obtain the search feature f_s and language feature f_l, respectively. Subsequently, in section 3.3, we utilize a prompt modulation module to filter out the inconsistent description in the initial verbal reference and the visual reference, thus forming more precise prompt information to guide the target location. Finally, in section 3.4, a target decoding module is responsible for integrating the multi-modal prompts and performing target retrieval within the search image.

3.2. Feature Extraction and Enhancement

Visual backbone. Considering the notable achievements of transformer models in image processing, we adopt the vanilla Swin-Transformer [21] as our visual backbone. To strike a balance between tracking accuracy and computational cost, we retain only the first three stages of the Swin-Transformer architecture, with the output of the third stage serving as our visual feature representation. For the input search image $I_s \in \mathbb{R}^{3\times H_s \times W_s}$, we feed it into the visual backbone and a channel adjustment layer to obtain the search region feature $\boldsymbol{f}_s \in \mathbb{R}^{N_s \times C}$, where N_s and C denote the number of features and channels, respectively. Herein we set $C = 256$.

Text backbone. To process the language description $\mathcal{D}$, we employ the widely adopted linguistic embedding model, RoBERTa [20], for the extraction of textual features. A projection layer is added behind the text backbone for adjusting feature dimensions. The output text feature is denoted as $\boldsymbol{f}_l \in \mathbb{R}^{L\times C}$, where L represents the length of the input text and $C = 256$ corresponds to the number of channels.

Feature enhancement. To extract discriminative features, we employ a bi-attention mechanism between the search image and the target reference for feature enhance-

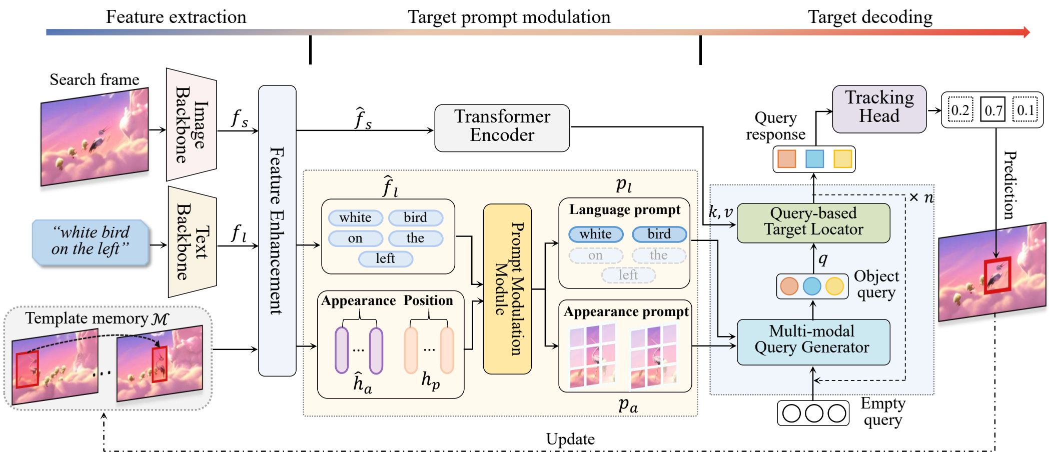


Figure 2. Overview of our proposed framework. It comprises three key components: a feature extraction module for extracting image and text features, a prompt modulation module that generates precise appearance and language descriptions of the target, and a target decoding module that jointly establishes the correlation between the search image and the multi-modal target prompts for target retrieval.

ment. In detail, the search feature f_s attends to both the text feature f_l and the historical appearance feature h_a to obtain the enhanced search feature $\hat{f}_s$, while the text feature f_l and the historical appearance feature h_a separately attend to the search feature f_s to obtain the enhanced feature $\hat{f}_l$ and $\hat{h}_a$. This process can be formulated as:

$$\hat{f}_s = \omega_s(f_s + \text{softmax}(\frac{f_s f_l^T}{\sqrt{C}}) f_l + \text{softmax}(\frac{f_s h_a^T}{\sqrt{C}}) h_a), \tag{1}$$

$$\hat{f}_l = \omega_l(f_l + \text{softmax}(\frac{f_l f_s^T}{\sqrt{C}}) f_s), \tag{2}$$

$$\hat{h}_a = \omega_a(h_a + \text{softmax}(\frac{h_a f_s^T}{\sqrt{C}}) f_s), \tag{3}$$

where ω_s, ω_l and ω_a are linear layers. To avoid disturbing the motion information of the object, here we only augment the appearance feature in template memory.

3.3. Target Prompt Modulation

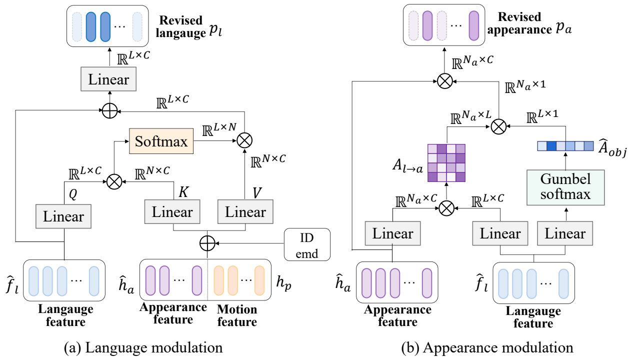
Accurate target cue information is essential for target tracking. However, due to the dynamics of the target in the course of tracking, the state description in language may not match the current target. As depicted in Fig. 2, the positional state "on the left" in the language description corresponds to the object in the initial frame. However, as the object moves in subsequent frames, this description no longer aligns with the object. In fact, the object is currently in the middle of the image. Meanwhile, due to mutual occlusion between objects, the object's appearance features in the template memory may include background features. For instance, in Fig. 2, the red bounding box erroneously encompasses the yellow bird, bringing further interference to the tracker. Using inaccurate language features and appearance features as the object prompt to retrieve the target in the search region may lead to tracking drift. To address this issue, we present a multi-modal prompt modulation module that exploits the complementarity between dynamic historical information and the language description, facilitating the formation of a more accurate target prompt.

Language modulation. We use motion cues from the template memory to adjust the language description. Specifically, h_p in the $\mathcal{M}$ stores the object position information of multiple previous frames, serving as a motion cue to assess whether the state description in the text feature $\hat{f}_l$ aligns with the current scene. Meanwhile, $\hat{h}_a$ in the $\mathcal{M}$ contains the object appearance information, acting as a visual cue to evaluate whether the target appearance description in the text feature $\hat{f}_l$ is correct. As shown in Fig. 3(a), we utilize a multi-head cross-attention operation (MHCA) to generate the language prompt. Before inputting to a cross-attention network, we first apply a self-attention operation on the $\hat{h}_a$ and h_p, respectively, to capture temporal changes in appearance and position. Subsequently, we add an appearance identifier vector $v_a \in \mathbb{R}^C$ to $\hat{h}_a$ and a motion identifier embedding $v_p \in \mathbb{R}^C$ to h_p. These two identifier vectors serve as indicators for different temporal cues. The process can be expressed by:

$$\hat{\mathcal{M}} = [\theta_a(\hat{h}_a) + [v_a]^{N_a}, \theta_p(h_p) + [v_p]^{N_p}], \tag{4}$$

where θ_a and θ_p represent a self-attention operation. $[\cdot]^n$ denotes the duplicate the vector n times and $[\cdot, \cdot]$ denotes the concatenation operation. N_a and N_p denotes the number of appearance feature and position feature saved in the template memory. The resulting $\hat{\mathcal{M}}$ is a matrix of size $N \times C$, where $N = N_a + N_p$.

Next, we utilize $\hat{f}_l$ as Query and linearly transform $\hat{\mathcal{M}}$ to obtain Key and Value for cross attention. This process can



(a) Language modulation

(b) Appearance modulation

Figure 3. Architecture of the proposed language prompt modulation module (a) and the appearance modulation module (b).

be formulated as:

$$p_l = \varphi_{agg}(\hat{f}_l + \text{softmax}(\frac{\varphi_q(\hat{f}_l)\varphi_k(\hat{\mathcal{M}})^T}{\sqrt{C}})\varphi_v(\hat{m})), \quad (5)$$

where $\varphi_{(\cdot)}$ represents different linear layer for feature transformation. After the above processing, the language features are re-weighted to generate the context-aware language prompt. Intuitively, the word that is more compatible with the template memory will be given higher attention, and the opposite will be given lower attention. We visualize the activation map of the language feature before and after modulation in Fig. 6(b).

Appearance modulation. The purpose of appearance modulation is to generate a binarized mask based on the category or appearance description of the tracked object in the sentence D, which can better fit the shape of the target. The appearance modulation is shown in Fig. 3(b). We first calculate a similarity matrix $A_{l\mapsto a}$ between $\hat{h}_a$ and $\hat{f}_l$:

$$A_{l\mapsto a} = \text{softmax}(\frac{\delta_a(\hat{h}_a)\delta_l(\hat{f}_l)^T}{\sqrt{C}}), \quad (6)$$

where δ_a and δ_l is a linear layer for feature transformation. This matrix $A_{l\mapsto a} \in \mathbb{R}^{N_a \times L}$ establishes the pixel-to-word correspondence, and pixels that correspond to the linguistic description yield high similarity scores. However, the tracking objects are often specified by describing their relative position to other objects, such as *"the fox on the bottom of the tree"*. In this case, the pixel belonging to the *"tree"* also gets a high similarity score. Therefore, we compute binarized subjecthood scores A_{obj} for the words via a Gumbel-Softmax [12, 22] operation. For the i^{th} word, its importance in the sentence is scored by:

$$A_{obj}^i = \frac{\exp(\mathbf{W}_{obj}\hat{f}_l^{\,i} + \gamma_i)}{\sum_{j=1}^{L}\exp(\mathbf{W}_{obj}\hat{f}_l + \gamma_j)} \quad (7)$$

where $\mathbf{W}_{obj} \in \mathbb{R}^{1\times C}$ is the weights of the learned linear projections for the text feature, γ_i and γ_j are random samples drawn from the Gumbel (0, 1) distribution. Then a subjecthood matrix score $\hat{A}_{obj} \in \mathbb{R}^{L\times 1}$ assigned all the words in the sentence is calculated by taking the one-hot [32] operation:

$$\hat{A}_{obj} = \text{one} - \text{hot}(A_{obj}^{argmax}) + A_{obj} - \text{sg}(A_{obj}), \quad (8)$$

where sg is the stop gradient operator. Finally, we multiply the $A_{l\mapsto a}$ and $\hat{A}_{obj}$ as the target mask, and the appearance prompt is formed by:

$$M_{obj} = A_{l\mapsto a} \times \hat{A}_{obj}, \quad (9)$$

$$p_a = \hat{h}_a \times M_{obj}, \quad (10)$$

where $M_{obj} \in \mathbb{R}^{N_a\times 1}$ indicates the probability that pixel belongs to the target. With such a design, we can filter out the background feature in the template, which produces a more accurate appearance prompt of the target for subsequent retrieval of the target.

3.4. Target Decoding

Appearance information and language information are both important for accurate object tracking. Different from the previous works, we treat language-based matching and appearance-based matching as a unified instance tracking problem and propose a target decoding module to achieve it. This module is composed of a query generator that aims to produce a query vector derived from the language prompt and appearance prompt. A query-based target locator that establishes the correlation between the search image and the query vector. The target decoding module is implemented by a transformer-based architecture proposed by a one-stage Deformable-DETR [42] for its flexible query-to-instance fashion.

As illustrated in Fig 4, an empty vector q_{init} is concurrently injected into the prompt information through an attention-based mechanism. Then it transforms into an object-aware query vector q_{obj}. The process can be formulated as:

$$q_{obj} = \psi(q_{init} + \alpha\sum_{i=0}^{N_t} a^i p_a^i + (1-\alpha)\sum_{j=0}^{N_l} b^j p_l^j), \quad (11)$$

where ψ is a linear layer, and α is a coefficient balancing the information from the different modal prompts. a^i and b^j denote the attention weights assigned to i^{th} element in appearance prompt and j^{th} element in the language prompt, respectively. Take the a^i for example, it is calculated by:

$$a^i = \frac{\exp(\mathbf{W}_a p_a^i \mathbf{W}_q q_{init})}{\sum_{k=0}^{N_t}\exp(\mathbf{W}_a p_a^k \mathbf{W}_q q_{init})}, \quad (12)$$

where $\mathbf{W}_a \in \mathbb{R}^{1\times C}$, $\mathbf{W}_q \in \mathbb{R}^{1\times C}$ represent a transformation matrix. When there is only language description as the target cue in the first frame of the sequence, we set α to 0 and the rest of the stage to a learnable parameter. In this way, our decoder can use the same parameters for single-modal or multi-modal target tracking. Afterward, in the query-based target locator, the object query q_{obj} performed cross-attention with the search image to infer the target. In detail, the search feature $\hat{f}_s$ is fed into the transformer en-

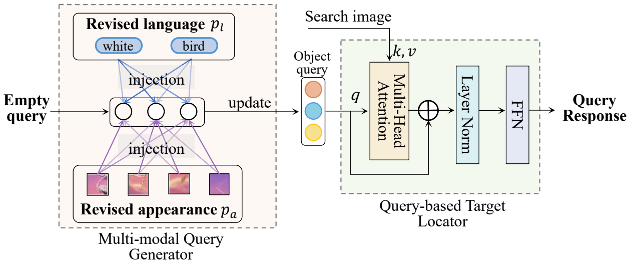


Figure 4. Architecture of the proposed target decoding module.

Table 1. Ablation study on the components of QueryNLT. All models are trained on the same training set and evaluated on TNL2K [30] under the *"NL"* setting.

#ID	Model	*AUC*
0	QueryNLT(Full Model)	53.3
1	w/o language modulation	52.2
2	w/o appearance modulation	51.6
3	separate matching	49.8
4	static template	51.0

Table 2. Ablation study on the query number on TNL2K [30] under the *"NL"* setting.

Query number	1	3	5	7
AUC	51.6	52.8	53.3	53.4

coder network to serve as the Key and Value for the cross attention, with the query vector q_{obj} acting as the Query. The object query is iteratively refined over stacked decoder layers and outputs a query response r which contains the target state within the search region.

To accommodate the diversity of linguistic expressions and target categories, which may result in a diversity of prompts, we adopt multiple vectors as a query set for target retrieval. Each query vector captures a unique interpretation of linguistic expressions and visual templates, emphasizing different aspects of the target. We conduct an ablation study on the number of queries as presented in Tab. 2.

Finally, we employ a classification head and a regression head to predict a score and box for each query. The bounding box with the highest score is selected as the final prediction. Following [42], the regression head is supervised by the L1 loss and GIoU [26] loss, and the classification head is supervised by cross-entropy loss.

After obtaining the position of the target in the current frame, we update the prediction result to template memory. In detail, the target feature is generated by a ROI align operation and is updated h_a. The corresponding query response which embeds the object's center position, width, and height in the current frame, is stored in h_p. Dynamically updated template memory facilitates the perception of the target's temporal changes.

4. Experiment

4.1. Implementation Details

The proposed QueryNLT is implemented in Pytorch on 6 NVIDIA RTX-3090 GPUs. We utilize Swin-B [21] pretrained on ImageNet [13] as the visual backbone. The RoBERTa [20] model is selected as the text backbone, with its parameters frozen throughout the entire training phase. The size of the template image and search image are set to 128 × 128 and 320 × 320, respectively. Following the training strategies of [41], our training dataset comprises TNL2K [30], OTB-Lang [18], LaSOT [4] and RefCOCOg[23], with an equal sampling ratio across the datasets. The batch size is set to 12 per GPU, with a total of 300 epochs. We implement a warmup strategy where the initial learning rates for the visual and other parameters increase linearly to 10^{-5} and 10^{-4}, respectively, within the first 30 epochs. Subsequently, the learning rates are reduced by a factor of 10 on the 200-th and 290-th epochs.

Following the protocols in [30], we evaluate our approach with two settings: (1) *"NL"*: the tracker is initialized with the natural language; (2) *"NL+BB"*: the tracker is initialized with both the natural language and the bounding box. To ensure consistency in training and inference, we extract three frames from each video for training our network. During this process, the tracker is first initialized using linguistic descriptions. Subsequently, visual templates stored in memory are jointly utilized to predict the target in the subsequent frames. During the inference phase, We update the template memory by replacing outdated trajectories with new ones, limiting the memory to a maximum capacity of three frames.

4.2. Ablation Study

To assess the effectiveness of our proposed components, we conduct an ablation study on the TNL2K [30] dataset under the *"NL"* evaluation setting. All variants are trained with the same training strategy as the full model. Tab. 1 reveals the significance of each component. Line 2 and line 3 show that accurate target reference is important to improve the discrimination of the tracker. When replaced with the modulated language feature and modulated appearance, the AUC score improved from 52.2% and 51.6% to 53.3%, presenting improvements of 1.1% and 1.7%, respectively. Model 3 is a variant that disentangles the language-search matching and template-search matching within the proposed framework. The prediction with the highest score of the query set is considered the target prediction. When replaced with the full model that simultaneously establishes the correspondence between multi-modal reference with search image, there is a notable improvement of 2.5% in AUC. This result demonstrates that the complementary nature of multi-modal information effectively boosts the holis-

Table 3. Comparison of our method with state-of-the-art approaches on OTB-Lang [18, 31], LaSOT [4] and TNL-2K [30] datasets. Top-2 results are highlighted in red and blue respectively.

Tracker	Initialize	OTB-Lang			LaSOT			TNL-2K		
		AUC	$Prec$	$NPrec$	AUC	$Prec$	$NPrec$	AUC	$Prec$	$NPrec$
SiamRPN++ [15]	BB	-	-	-	49.6	49.1	56.9	41.3	41.2	48.0
Ocean [38]	BB	-	-	-	56.0	56.6	65.1	38.4	37.7	45.0
AutoMatch [39]	BB	-	-	-	58.3	59.9	67.4	47.2	43.5	-
TrDiMP [28]	BB	-	-	-	63.9	61.4	-	52.3	52.8	-
TransT [1]	BB	-	-	-	64.9	69.0	73.8	50.7	51.7	-
SwinTrack-B [19]	BB	-	-	-	61.3	76.5	-	-	55.9	57.1
OSTrack-384 [36]	BB	-	-	-	71.1	77.6	81.1	55.9	-	-
TNLS-II [18]	NL	25.0	29.0	-	-	-	-	-	-	-
RVTNLN [5]	NL	54.0	56.0	-	-	-	-	-	-	-
RTTNLD [6]	NL	54.0	78.0	-	28.0	28.0	-	-	-	-
GTI [35]	NL	58.1	73.2	-	47.8	47.6	-	-	-	-
TNL2K-1 [30]	NL	19.0	24.0	-	51.1	49.3	-	11.4	6.4	11.0
CTRNLT [17]	NL	53.0	72.0	-	52.0	51.0	-	14.0	9.0	-
JointNLT [41]	NL	59.2	77.6	-	56.9	59.3	64.5	54.6	55.0	70.6
JointNLT [41] *	NL	57.8	77.0	70.5	52.8	54.4	60.8	52.1	51.2	68.8
Ours	NL	61.2	81.0	73.9	54.2	55.0	62.5	53.3	53.0	70.4
TNLS-III [18]	NL+BB	55.0	72.0	-	-	-	-	-	-	-
RVTNLN [5]	NL+BB	67.0	73.0	-	50.0	56.0	-	25.0	27.0	34.0
RTTNLD [6]	NL+BB	61.0	79.0	-	35.0	35.0	-	25.0	27.0	33.0
SNLT [7]	NL+BB	66.6	80.4	-	54.0	57.6	-	27.6	41.9	-
TNL2K-2 [30]	NL+BB	68.0	88.0	-	51.0	55.0	-	41.7	42.0	50.0
JointNLT [41]	NL+BB	65.3	85.6	79.5	60.4	63.6	69.4	56.9	58.1	73.6
JointNLT [41] *	NL+BB	63.6	87.1	78.8	58.8	62.3	68.7	56.6	57.9	74.8
Ours	NL+BB	66.7	88.2	82.4	59.9	63.5	69.6	57.8	58.7	75.6

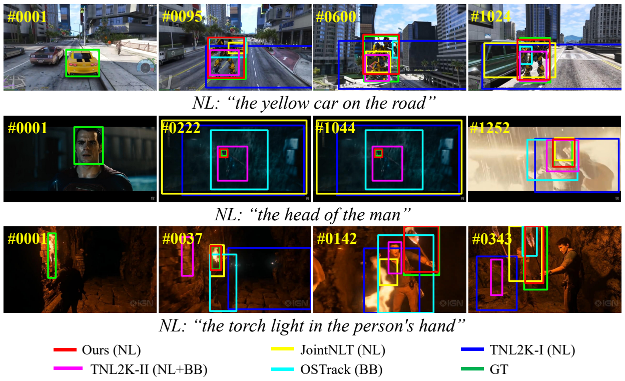
* our reproducing results using the officially released code.

tic understanding and perception of the target. Model 4, relying solely on grounded results as the visual template, achieved an AUC score of 51.0%. However, introducing multiple dynamic templates led to a significant improvement from 51.0% to 53.3%, underscoring the crucial role of temporal information in enhancing tracking robustness.

We provide an analysis of the impact of the query number for each frame as shown in Tab. 2. The model consistently demonstrates significant results across all settings. Overall, the performance improves with an increase in the query count. There is no noticeable improvement when the query number increases from 5 to 7. This observation suggests that a count of 5 already provides a sufficient variety of combinations to comprehend the cues. It is noteworthy that since queries are processed in parallel, an increase in the number of queries does not affect the speed of the tracker.

4.3. State-of-the-art Comparison

In this section, we compare our approach with state-of-the-art trackers, including JointNLT [41], CTRNLT [17], TNL2K [30] and others approaches, on three challenging natural language tracking datasets. Following the protocols in [30], we test the performance of our approach for initialization using only *"NL"* and using *"NL+BB"*. We also provide experimental results on the visual grounding dataset to demonstrate that our proposed method is effective in establishing text-image correlation. All comparison results are obtained from the paper. Additionally, we retrained the JointNLT [41] using the official release code, denoted as "JointNLT*". By deploying in the same experimental setting, the result of "JointNLT*" serves as an important baseline to measure the effectiveness of our method.



NL: "the yellow car on the road"

NL: "the head of the man"

NL: "the torch light in the person's hand"

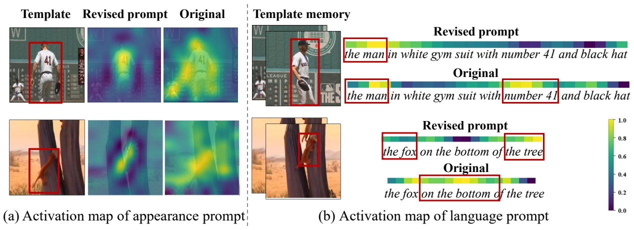
Figure 5. Qualitative comparisons of the proposed QueryNLT with the state-of-the-art trackers on three challenging sequences. Our QueryNLT can accurately target locations even when objects suffer from severe appearance variations, background clutters, and similar distractors.

Table 4. Comparison of our method with state-of-the-art approaches for visual grounding on RefCOCOg [23] dataset.

Method	LBYL [11]	ReSC [34]	TransVG [3]	VLTVG [33]	JointNLT [41]	Ours
val-g	62.70	63.12	67.02	73.0	70.07	72.0
val-u	-	67.3	68.67	76.0	-	75.3
test-u	-	67.2	67.73	74.2	-	73.2

Evaluation on TNL2K dataset. TNL2K is a benchmark specifically designed for evaluating natural language-guided tracking algorithms. It comprises a diverse collection of videos, including natural, animation, infrared, and virtual game videos, thereby facilitating a comprehensive evaluation of the framework's adaptability across different domains. The rich and discriminative annotated language makes the TNL2K dataset particularly well-suited for the task of tracking based solely on natural language descriptions. As shown in Tab. 3, under the *"NL"*, our QueryNLT is the second best only behind the JointNLT [41] but surpasses the reproduction of the model JointNLT* by 1.2%, 1.8% and 1.6% on three metrics. Under the *"NL+BB"* setting, our QueryNLT performs best in terms of all indicators. Compared with TNL2K-2 [30], which employs adaptive switching between templates and language cues for target inference, our proposed approach (NL+BB) achieves notable improvements of 15.8%, 16.7%, and 25.1% in terms of AUC, precision, and normalized precision, respectively. It demonstrates the complementary nature of multi-modal information in recognizing targets. Besides, our QueryNLT outperforms JointNLT [41], which utilizes a static language description across all video frames, achieving superiority by 0.9%, 0.6%, and 2.0% in terms of three metrics. The result emphasizes the effectiveness of dynamic and context-aware linguistic descriptions for improving tracking performance. Qualitative results are provided in Fig. 5.

Evaluation on OTB-Lang dataset. The OTB-Lang dataset is originally released in [31] and later extended with a sentence description of the target object per video by [18]. It encompasses 11 challenging interference attributes, such as motion blur, scale variation, occlusion, out-of-view scenarios, background clutter, and more. The results on OTB-Lang are shown in Tab. 3. Remarkably, our proposed approach outperforms all other trackers under the *"NL"* setting. Specifically, our proposed QueryNLT surpasses JointNLT [41] by 2.0% and 3.4% in terms of AUC and precision, respectively. And compared with JointNLT*, our approach shows improvements of 3.4%, 4.0%, and 3.4% in three metrics. Additionally, under the *"NL+BB"* setting, our QueryNLT is the second best in terms of AUC, only behind TNL2K [30], within which the tracking module is trained on a larger training dataset. These results collectively highlight the robustness of our proposed approach, indicating its ability to effectively handle various challenging factors encountered in tracking tasks.



(a) Activation map of appearance prompt (b) Activation map of language prompt

Figure 6. Visualization of the appearance and language prompts.

Evaluation on LaSOT dataset. The LaSOT is a long-term tracking dataset that provides both bounding box and natural language annotations. It comprises 1120 training video sequences and 280 testing video sequences. It should be noted that the linguistic information in LaSOT lacks a description of the relative positions of the objects, and thus the given linguistic description is ambiguous when similar objects are interfering. This means that this dataset is not suitable for accomplishing language-assist tracking tasks, and a similar view can be found in [30, 35]. Here we mainly discuss the comparison results under the *"NL+BB"* setting. As shown in Tab. 3, our proposed approach achieves the performance of 57.8% 58.7%, and 75.6% in terms of AUC, precision, and normalization precision, respectively. It surpasses the TNL2K-2 [30] by 5.9% in AUC and 5.9% in precision. These results demonstrate that our approach is competitive for long-term tracking tasks.

Evaluation on RefCOCOg dataset. We evaluate the visual grounding performance on both the validation and test sets of the RefCOCOg dataset [23]. The assessment is conducted using the average IoU as the evaluation metric." As shown in Tab. 4, our method is second only to VLTVG [33] although we are not specialized for visual grounding. This also explains that our tracking method can perform robust tracking even when only the language description is given.

5. Conclusion

In this paper, we have introduced a unified framework for natural language tracking that effectively leverages both visual and verbal references to improve target perception and discrimination. We proposed the prompt modulation module to filter out the description in target references, thus forming accurate and context-aware visual and verbal cues. Besides, the target decoding module is designed to integrate multi-modal reference information to reason about the position of the target within the search image. Incorporating the target decoding network with precise target prompts greatly improves the discrimination of the tracker. Extensive experiments on the natural language tracking datasets and the visual grounding dataset demonstrate our proposed approach achieves competitive performance.

References

[1] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8126–8135, 2021. 7

[2] Ying Cui, Dongyan Guo, Yanyan Shao, Zhenhua Wang, Chunhua Shen, Liyan Zhang, and Shengyong Chen. Joint classification and regression for visual tracking with fully convolutional siamese networks. *Int. J. Comput. Vis.*, pages 1–17, 2022. 1

[3] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Int. Conf. Comput. Vis.*, pages 1769–1779, 2021. 8

[4] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2, 6, 7

[5] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Robust visual object tracking with natural language region proposal network. *arXiv preprint arXiv:1912.02048*, 1(7):8, 2019. 1, 7

[6] Qi Feng, Vitaly Ablavsky, Qinxun Bai, Guorong Li, and Stan Sclaroff. Real-time visual object tracking with natural language description. In *IEEE Conf. Appli. Comput. Vis.*, pages 700–709, 2020. 1, 2, 3, 7

[7] Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5851–5860, 2021. 3, 7

[8] Maximilian Filtenborg, Efstratios Gavves, and Deepak Gupta. Siamese tracking with lingual object constraints. *arXiv preprint arXiv:2011.11721*, 2020. 3

[9] Dongyan Guo, Yanyan Shao, Ying Cui, Zhenhua Wang, Liyan Zhang, and Chunhua Shen. Graph attention tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9543–9552, 2021. 1

[10] Mingzhe Guo, Zhipeng Zhang, Heng Fan, and Liping Jing. Divert more attention to vision-language tracking. *Adv. Neural Inform. Process. Syst.*, 35:4446–4460, 2022. 3

[11] Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. Look before you leap: Learning landmark features for one-stage visual grounding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16888–16897, 2021. 8

[12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 5

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Comm. of the ACM*, 60(6):84–90, 2017. 6

[14] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8971–8980, 2018. 3

[15] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4282–4291, 2019. 1, 3, 7

[16] Xin Li, Yuqing Huang, Zhenyu He, Yaowei Wang, Huchuan Lu, and Ming-Hsuan Yang. Citetracker: Correlating image and text for visual tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9974–9983, 2023. 3

[17] Yihao Li, Jun Yu, Zhongpeng Cai, and Yuwen Pan. Cross-modal target retrieval for tracking by natural language. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4931–4940, 2022. 7

[18] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6495–6503, 2017. 1, 2, 3, 6, 7, 8

[19] Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. *Adv. Neural Inform. Process. Syst.*, 35:16743–16754, 2022. 7

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3, 6

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, pages 10012–10022, 2021. 3, 6

[22] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 5

[23] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11–20, 2016. 2, 6, 8

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021. 3

[25] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In *AAAI*, pages 4780–4789, 2019. 3

[26] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 658–666, 2019. 6

[27] Yanyan Shao, Qi Ye, Wenhan Luo, Kaihao Zhang, and Jiming Chen. Intertracker: Discovering and tracking general objects interacting with hands in the wild. In *Int. Conf. on Intell. Robots and Systems*, pages 9079–9085. IEEE, 2023. 3

[28] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1571–1580, 2021. 7

[29] Xiao Wang, Chenglong Li, Rui Yang, Tianzhu Zhang, Jin Tang, and Bin Luo. Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking. *arXiv preprint arXiv:1811.10014*, 2018. 3

[30] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13763–13773, 2021. 1, 2, 3, 6, 7, 8

[31] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37 (9):1834–1848, 2015. 2, 7, 8

[32] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18134–18144, 2022. 5

[33] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9499–9508, 2022. 8

[34] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Eur. Conf. Comput. Vis.*, pages 387–404, 2020. 8

[35] Zhengyuan Yang, Tushar Kumar, Tianlang Chen, Jingsong Su, and Jiebo Luo. Grounding-tracking-integration. *IEEE Trans. Circuit Syst. Video Technol.*, 31(9):3433–3443, 2020. 1, 2, 7, 8

[36] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Eur. Conf. Comput. Vis.*, pages 341–357, 2022. 7

[37] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Eur. Conf. Comput. Vis.*, pages 341–357. Springer, 2022. 1

[38] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *Eur. Conf. Comput. Vis.*, pages 771–787, 2020. 7

[39] Zhipeng Zhang, Yihao Liu, Xiao Wang, Bing Li, and Weiming Hu. Learn to match: Automatic matching network design for visual tracking. In *Int. Conf. Comput. Vis.*, pages 13339–13348, 2021. 7

[40] Chongyang Zhao, Yuankai Qi, and Qi Wu. Mind the gap: Improving success rate of vision-and-language navigation by revisiting oracle success routes. In *ACM Int. Conf. Multimedia*, pages 4349–4358, 2023. 3

[41] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 23151–23160, 2023. 3, 6, 7, 8

[42] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5, 6

[43] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *Int. Conf. Learn. Represent.*, 2017. 3