Diffusion Mask-Driven Visual-language Tracking

Guangtong Zhang[1,2], **Bineng Zhong** *[1,2], **Qihua Liang**[1,2], **Zhiyi Mo**[1,2,3], **Shuxiang Song**[1,2]
[1]Key Laboratory of Education Blockchain and Intelligent Technology Ministry of Education, Guangxi Normal University, Guilin 541004, China.
[2]Guangxi Key Lab of Multi-Source Information Mining & Security, Guangxi Normal University, Guilin 541004, China.
[3]Guangxi Key Laboratory of Machine Vision and Intelligent Control, Wuzhou University,Wuzhou 543002, China.
zhangguangtong@stu.gxnu.edu.cn, {bnzhong,qhliang}@gxnu.edu.cn, zhiyim@gxuwz.edu.cn, songshuxiang@mailbox.gxnu.edu.cn

Abstract

Most existing visual-language trackers greatly rely on the initial language descriptions on a target object to extract their multi-modal features. However, the initial language descriptions are often inaccurate in a highly time-varying video sequence and thus greatly deteriorate their tracking performance due to the low quality of extracted multi-modal features. To address this challenge, we propose a Diffusion Mask-Driven Visual-language Tracker (DMTrack) based on a diffusion model. Confronting the issue of low-quality multi-modal features due to inaccurate language descriptions, we leverage the diffusion model to capture high-quality semantic information from multi-modal features and transform it into target mask features. During the training phase, we further enhance the diffusion model's perception of pixel-level features by calculating the loss between the target mask features and the ground truth masks. Additionally, we perform joint localization of the target using both target mask features and visual features, instead of relying solely on multi-modal features for localization. Through extensive experiments on four tracking benchmarks (i.e., LaSOT, TNL2K, LaSOT$_{ext}$, and OTB-Lang), we validate that our proposed Diffusion Mask-Driven Visual-language Tracker can improve the robustness and effectiveness of the model.

1 Introduction

In recent years, the integration of computer vision and natural language processing has received extensive attention from researchers[Alec *et al.*, 2021], providing an opportunity for innovative research in visual and textual information. Describing the state features and future trends of the targets in language allows the trackers to be more robust in complex scenarios[Zheng *et al.*, 2023]. However, the existing visual-language tasks annotate language descriptions based on the state and trend of the target in the first frame. This annotation method leads to inaccuracies and even side effects in language descriptions as the target moves, resulting in a decrease in the quality of multi-modal features. Current visual-language trackers, by directly utilizing multi-modal features for target localization, face tracking failures due to the low quality of these multi-modal features. Therefore, capturing higher-quality semantic information from the existing multi-modal information becomes a challenge.

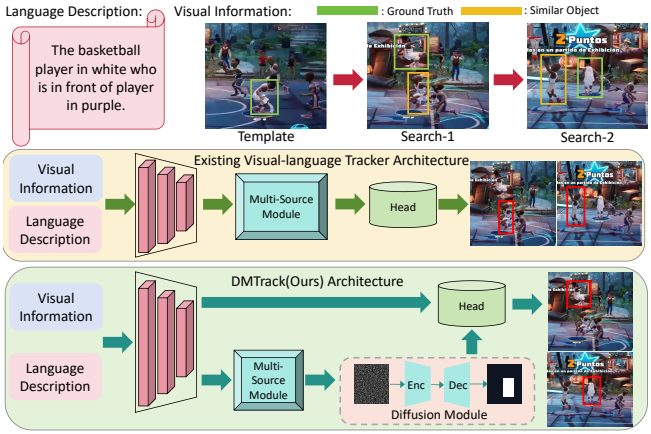

Figure 1: The comparison of the difference architectures between our DMTrack and existing visual-language trackers[Li *et al.*, 2023] . Our DMTrack exhibits enhanced robustness in a scenario where the language information is inaccurate.

It is well-known that the diffusion models excelled in image generation tasks[Ho *et al.*, 2020] have recently demonstrated outstanding capabilities in dense prediction tasks. Jordan et al.[Tomer *et al.*, 2021] first introduce diffusion models in the field of image segmentation and Dmitry et al.[Dmitry *et al.*, 2022] prove that diffusion models can effectively capture semantic information from input images, providing more efficient pixel-level representations for segmentation models. In camouflage object detection tasks, Chen et al.[Chen *et al.*, 2024] surpass the best performance in existing camouflage object detection tasks by introducing diffusion models. In-

*Bineng Zhong is the corresponding author.

spired by the success of the diffusion models, we integrate a diffusion model into the visual-language tracking task. However, the current application of the diffusion model is mainly focused on dense prediction tasks. Applying the diffusion model to visual-language tracking tasks still faces numerous challenges. In dense prediction tasks, models often need to identify different categories of targets without the necessity of distinguishing among objects of the same class. In visual-language tracking, however, the models need to distinguish among objects of the same class to accurately locate a target in a scene with similar object interference. To address these challenges, we propose a feature filtering module, double sample training method, and candidate elimination module for diffusion ground truth to adapt diffusion models to the visual-language tracking task.

By successfully integrating a diffusion model, DMTrack addresses the issue of inaccurate initial language descriptions caused by highly time-varying target motion and appearance variations in a video sequence. DMTrack leverages the diffusion model to capture high-quality semantic information from multi-modal features and transforms it into target mask features. During the training phase, DMTrack further enhances the diffusion model's perception of pixel-level features by calculating the loss between the target mask features and the ground truth masks. In the testing phase, the generated target mask features by the diffusion model undergo multiple diffusion samplings to generate more accurate target mask features. Finally, joint localization of the target is performed using both target mask features and visual features, rather than relying solely on multi-modal features for localization. The differences in architecture between DMTrack and existing visual-language trackers are illustrated in Fig.1. After the object moves, in Search-1, two targets match the language description, while in Search-2, no target matches the language description. Compared to existing structure trackers, DMTrack demonstrates greater robustness in scenarios where language descriptions are inaccurate. In summary, our contributions are as follows:

- We propose a visual-language tracker based on diffusion models. This tracker utilizes diffusion models to capture semantic information in visual-language features and provide pixel-level target information to the model.
- To our knowledge, we are the first to apply diffusion models to visual-language tracking tasks and propose a feature filtering module, double sample training method, and candidate elimination module for diffusion ground truth to adapt diffusion models to the visual-language tracking task.
- We achieve excellent algorithm performance on multiple visual-language tracking datasets to demonstrate the superiority of our proposed tracker.

2 Related Work

2.1 Visual-language Tracking

With the rapid development of visual trackers[Chen *et al.*, 2022; Cui *et al.*, 2022], visual-language tracking[Qi *et al.*, 2021] is also receiving increasing attention. Li et al.[Zhenyang *et al.*, 2017a] spearhead the amalgamation of language and vision for concurrent tracking endeavors, introducing a language-specific feature extraction network that acted as a catalyst for subsequent advancements in visual-language tracking. Yang et al.[Yang *et al.*, 2020] taxonomic visual-language tracking into three distinct subtasks—namely, grounding, tracking, and integration—proposing three distinct modules to address each subtask in isolation. Feng et al.[Qi *et al.*, 2021] propose a detection-tracking paradigm that employs language descriptions to furnish comprehensive tracking recommendations for the target in each frame. Wang et al.[Xiao *et al.*, 2021] aiming to standardize natural language tracking methodologies, introduced a novel benchmark for visual-language tracking denominated TNL2k. They also propounded two foundational approaches, initialized respectively by natural language and bounding boxes. Li et al.[Yihao *et al.*, 2022] introduce a target retrieval module for tracking, seamlessly incorporating it into a localized tracking apparatus. Guo et al.[Guo *et al.*, 2022] devise an asymmetric model structure leveraging language for the selection of visual information. Zhou et al.[Li *et al.*, 2023] devise a unified visual-language grounding and tracking framework grounded in language descriptions for the localization of reference objects. Meanwhile, Zheng et al.[Zheng *et al.*, 2023] serialize language descriptions and bounding box sequences into a sequence of discrete tokens, directly forecasting the spatial coordinates of the target in an auto-regressive fashion. Despite the achievements of these tracking methods, inaccurate language descriptions due to target motion affect the quality of multi-modal features. Existing visual-language trackers, relying on direct usage of low-quality multi-modal features, face tracking failures. In response to this challenge, DMTrack utilizes the diffusion model to capture high-quality semantic information from multi-modal features, addressing the issue of inaccurate language descriptions.

2.2 Diffusion Model

The diffusion model has garnered considerable attention in recent years as a technology with promising applications. Leveraging a parameterized Markov chain, these model[Ho *et al.*, 2020; Song *et al.*, 2020] undertake the denoising of data samples originating from random noise. Initially employed in the field of image without explicit ground truth, recent investigations[Zhang and Agrawala, 2023; Dhariwal and Nichol, 2021] have showcased its effectiveness in tackling practical challenges such as super-resolution[Li *et al.*, 2022; Wang *et al.*, 2021], deblurring[Whang *et al.*, 2022; Lee *et al.*, 2022], and image segmentation[Baranchuk *et al.*, 2021; Amit *et al.*, 2021]. Notably, the diffusion model exhibits distinctive potential in various segmentation tasks, with notable applications in remote sensing change detection[Bandara *et al.*, 2022] and medical image segmentation[Wolleb *et al.*, 2022]. For example, Wolleb et al.[Wolleb *et al.*, 2022] introduce a diffusion model for lesion segmentation employing images as prior information, while Rahman et al.[Rahman *et al.*, 2023] explore the diffusion model's capabilities in segmenting blurred images. However, the diffusion model is primarily applied in dense prediction tasks. Due to signif-

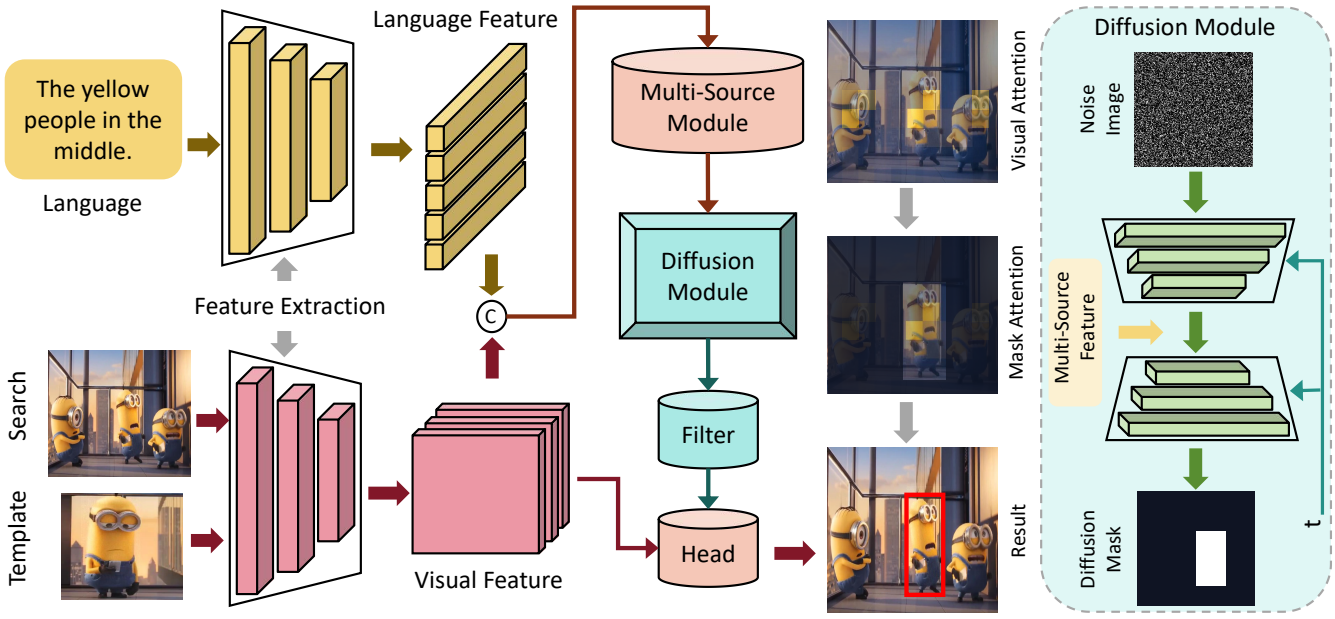

Figure 2: The overall framework of DMTrack. DMTrack extracts modal features through a feature extraction network and performs multi-source interaction on these features through a multi-source module. Using the multi-source features as conditions, the diffusion model is employed to restore the noisy image to the target position mask. The optimized target position mask features, enhanced by the feature filtering module, are combined with visual features to locate the target.

icant differences between dense prediction tasks and visual-language tracking tasks, there is a need for further exploration on how to integrate the diffusion model into visual-language tracking. To address this challenge, DMTrack has designed a training approach involving double sampling, a feature filtering module, and a ground truth candidate elimination module. These components aim to facilitate the adaptation of the diffusion model to visual-language tracking tasks.

3 Method

The overall framework of DMTrack proposed by us is illustrated in Fig.2. DMTrack is composed of a Visual and Language Feature Extraction Module, a Multi-modal Interaction Module, a Diffusion Module, and Head and Loss.

3.1 Visual and Language Feature Extraction.

Visual Feature Extraction. We chose the simple and efficient ViT model as the visual encoder of our DMTrack and pre-train it using OSTrack. The input to the visual encoder includes a pair of images, namely the template frame $z \in \mathbb{R}^{3\times H_z\times W_z}$, the search frame $x \in \mathbb{R}^{3\times H_x\times W_x}$. The template frame and search frame are first split and flattened into sequences of patches $z_p \in \mathbb{R}^{N_z\times\left(3\cdot P^2\right)}$ and $x_p \in \mathbb{R}^{N_x\times\left(3\cdot P^2\right)}$, where $P\times P$ is the resolution of each patch and $N_z = H_zW_z/P^2, N_x = H_xW_x/P^2$ is the number of patches in the template frame and search frame, respectively. Subsequently, using a trainable linear projection layer, the patches z_p and x_p are mapped to a D-dimensional latent space, referred to as patch embeddings. The learnable position embedding is then incorporated into the patch embedding of the template frame $\boldsymbol{B}_z$ and the search frame $\boldsymbol{B}_x$, generating the final template token embedding $\boldsymbol{H}_z^0 \in \mathbb{R}^{N_z\times D}$ and search frame token embedding $\boldsymbol{H}_x^0 \in \mathbb{R}^{N_x\times D}$. This process can be represented as follows:

$$\begin{aligned} H_z^0 &= \left[z_p^1E; z_p^2E; \cdots; z_p^{N_z}E\right] + B_z, \\ &E \in \mathbb{R}^{\left(3\cdot P^2\right)\times D}, B_z \in \mathbb{R}^{N_z\times D}, \\ H_x^0 &= \left[x_p^1E; x_p^2E; \cdots; x_p^{Nx}E\right] + B_x, \\ &B_x \in \mathbb{R}^{N_x\times D}. \end{aligned} \tag{1}$$

Language Feature Extraction. We choose the classic language feature extraction model BERT[Jacob *et al.*, 2019] as the language encoder for our tracker. Given a natural language description as the query Q_l, we first tokenize the language description and inject CLS and SEP tokens, resulting in a token sequence $L = \{\mathrm{CLS}, \boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_N, \mathrm{SEP}\}$, where N is the maximum length of the language query. Then, we input the token sequence into our language encoder to obtain language token features $\boldsymbol{T}_q \in \mathbb{R}^{C_q\times N}$, where $C_q = 768$ represents the output embedding dimension.

Multi-modal Interaction Module. Past visual-language trackers typically design a complex visual-language interaction module for the interaction between visual features and language information. However, this interaction method increases the model's complexity. Given that transformers are capable of effectively modeling relationships between language and visual features, we employ a Gated Cross-Attention mechanism to achieve fine-grained cross-modal relationship modeling between visual features and language information. This process can be represented as follows:

$$F_s = CrossAttn\left(H_x; T_q\right) + H_x, \tag{2}$$

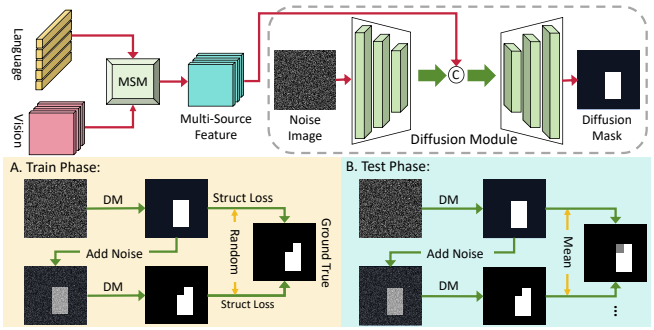

Figure 3: The diffusion process diagram used by DMTrack. During the training phase, we employ a double-sampling method and randomly select the sampled mask features. In the testing phase, we average the mask features obtained through multiple samplings. Where C represents feature concat.

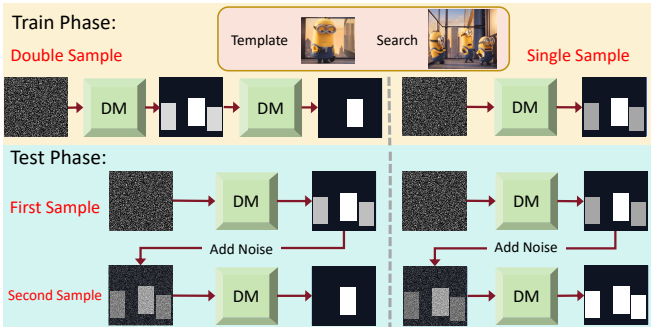

Figure 4: The schematic diagram of double sampling during the training phase. Compared to single sampling during the training phase, the double sampling training approach enables the diffusion model to better adapt to complex scenes.

where, H_x and T_q represent the outputs of the visual feature extraction module and the language feature extraction module, respectively.

3.2 Diffusion Module

Background. The diffusion mask-driven visual-language tracker proposed by us is based on a diffusion model, where the diffusion model transforms the noise $x_T \sim (0, I)$ into smaller noise samples x_t, converting the noise x_T into the sample x_0. We refer to this process as the forward diffusion process, which can be represented as:

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right), \quad (3)$$

where variance is controlled by noise schedule $\beta_t \in (0, 1)$. The marginal distribution of x_t can be directly obtained from the data x_0. This process can be described as:

$$q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right), \quad (4)$$

where $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_t, \alpha_t = 1 - \beta_t$. Starting from $p\left(\mathbf{x}_T\right) = \mathcal{N}\left(\mathbf{x}_T; \mathbf{0}, \mathbf{I}\right)$, the reverse diffusion process utilizes a U-shaped network f_θ to implement a series of incremental denoising steps, obtaining clean mask features. This network's reverse diffusion process can be described as:

$$p\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) := \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta\left(\mathbf{x}_t, t\right), \Sigma_\theta\left(\mathbf{x}_t, t\right)\right). \quad (5)$$

The entire process of the diffusion model is illustrated in Fig.3. The multi-source features obtained from the multi-source module serve as the diffusion conditions, allowing the restoration of the noisy image to the target's position mask. During the training phase, the diffusion model adopts a double-sampling approach. When calculating gradients, it randomly selects either the first diffusion-sampled image, the second sampled image, or the average of both, and calculates the loss against the ground truth. In the testing phase, the results of multiple diffusion samplings are averaged.

Double Sample Train. The current diffusion models employ a training approach where ground truth is added with noise as the initial noisy image for the diffusion model during the training phase. Single-sampling is then used to calculate the loss, and during the testing phase, a gradual noise reduction diffusion method is applied. However, this traditional diffusion model is not suitable for the visual-language tracking task. On one hand, the diffusion model lacks the capability to handle complex scenes, especially interference from similar objects. On the other hand, it tends to overly rely on the results of the previous sampling. The process of our proposed Double-Sampling method during the training phase is illustrated in Fig.4. On one hand, during the initial sampling in the training phase, the noisy image is generated as completely random noise and is not combined with the ground truth. This approach effectively suppresses the diffusion model's excessive reliance on the previous sampling, thereby improving tracking performance. On the other hand, we employ the double-sampling method to train the diffusion model during the training phase. The key to this method is that the noisy image for the second diffusion sampling is the result of the first diffusion sampling. This training approach provides the diffusion model with more complex scenarios, enhancing its robustness in handling complex scenes.

Ground Truth Candidate Elimination Module. The ground truth bounding box for object tracking not only contains the target but also includes a substantial amount of background information. This background information significantly impacts the diffusion effectiveness of the diffusion model and the precision of the pixel-level information provided to the model. Therefore, we propose a Ground Truth Candidate Elimination module to reduce the background information within the target box, as shown in Fig.5. In the early stages of the training phase, the ground truth values for the diffusion model are the ground truth bounding boxes in the tracking task. In the later stages of training, the introduced Ground Truth Candidate Elimination module is utilized to reduce background information within the ground truth. We have adopted the early elimination module from OSTrack[Ye *et al.*, 2022], and specific details can be referenced in the discussions on this module within OSTrack. However, unlike its original purpose of reducing computational load during the attention stage of the training phase, we use annotated background information from the network to eliminate background information within the ground truth. With the assistance of more accurate ground truth masks, the diffusion model can generate more precise mask features for noisy im-

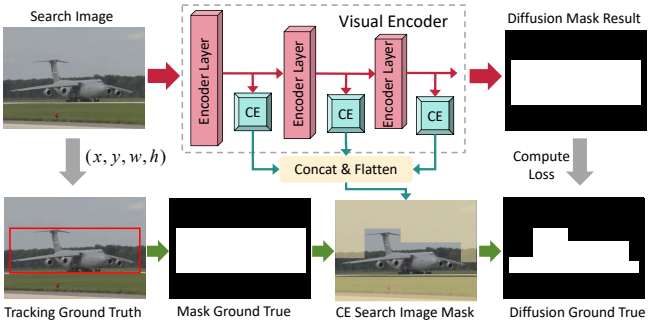

Figure 5: The flowchart of the ground truth candidate elimination module. We utilize the candidate elimination module to reduce the background information contained in the ground truth bounding box, enhancing the sampling effectiveness of the diffusion model.

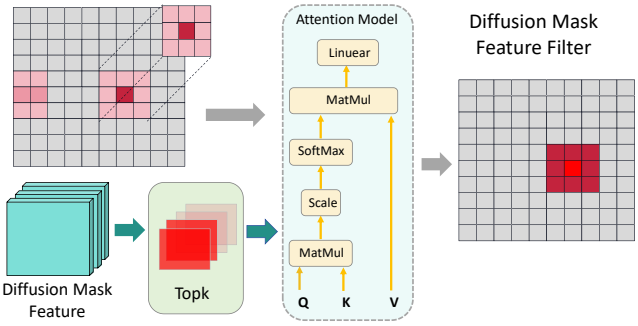

Figure 6: The schematic diagram of the feature filtering module. This module can enhance the response of important features while suppressing other interfering features.

ages, providing the model with more pixel-level information.

Feature Filtering Module. To enhance the quality of pixel-level mask features provided by the diffusion model for the tracker, we propose a Feature Filtering Module. The main idea of this module is to select and redistribute attention to features with the highest responses while suppressing other response features. The structure of the Feature Filtering Module is illustrated in Fig.6. This module filters the target position mask features obtained by the diffusion model to provide the tracker with higher-quality pixel-level mask features.

3.3 Head and Loss

Head. In the prediction head, we linearly combine the visual features with the diffused position mask. We employ a convolutional network composed of Conv-BN-ReLU layers for target classification and regression. The output of the convolutional network is considered as the score map $\boldsymbol{M} \in [0,1]^{\frac{H_x}{M} \times \frac{W_x}{M}}$ for object classification and the local offsets $\boldsymbol{D} \in [0,1]^{2 \times \frac{H_x}{M} \times \frac{W_x}{M}}$ to address the discretization errors caused by reduced resolution and normalized bounding box size (width and height) $\boldsymbol{S} \in [0,1)^{2 \times \frac{H_x}{M} \times \frac{W_x}{M}}$. In the object classification score map, the object position is determined as the location with the highest classification score, i.e., $(x_d, y_d) = \arg\max_{(x,y)} \boldsymbol{M}_{xy}$.

Loss. During the training process, we employed not only classification loss and regression loss but also utilized weighted focal loss for classification. Using the predicted bounding boxes, we performed bounding box regression using both ℓ_1 loss and IoU loss. In addition to these, we introduced structural loss for loss computation on the target location mask outputted by the diffusion model. Finally, the overall loss function is formulated as follows: The overall loss function can be formulated as:

$$L_{\text{track}} = L_{cls} + \lambda_{\text{iou}} L_{iou} + \lambda_{L_1} L_1 + \lambda_M L_{struct} \quad (6)$$

where $\lambda_{\text{iou}} = 2$, $\lambda_{L_1} = 5$ and $\lambda_M = 5$ are the regularization parameters in our experiments.

4 Experiments

4.1 Experimental Setup

We utilize the ViT[Dosovitskiy *et al.*, 2021] model pre-trained with OSTrack[Ye *et al.*, 2022] as our vision encoder and use the base uncased version of BERT[Jacob *et al.*, 2019] as our language encoder. The visual input to the DMTrack-256 network is an image pair consisting of a template patch of size 128×128, and a search patch of size 256×256. For the language input, the max length of the language is set to 36, including a CLS and a SEP token. We use the training splits of TNL2k[Xiao *et al.*, 2021], LaSOT[Fan *et al.*, 2018], OTB-Lang[Zhenyang *et al.*, 2017b], and RefCOCO-google[Junhua *et al.*, 2016] multiple training sets for joint training. Our model was implemented in the Pytorch framework on a server with 1 NVIDIA V100 GPU. Our model is trained with 100 epochs, each epoch with 60,000 image pairs and each mini-batch with 64 sample pairs. We also train the model using the AdamW optimizer, set the weight decay to 10^{-4}, the initial learning rate of the backbone to 2×10^{-5}, and other parameters to 2×10^{-4}. After 80 epochs, the learning rate is decreased by a factor of 10. We tested the proposed tracker on an NVIDIA 3080 GPU, and the single sample tracking speed is about 40 FPS.

4.2 Datasets and Metrics

In this section, we systematically assess the performance of the proposed model across four Visual-language (VL) tracking benchmarks, specifically TNL2k[Xiao *et al.*, 2021], LaSOT[Fan *et al.*, 2018], LaSOT$_{ext}$[Fan *et al.*, 2018] and OTB-Lang[Zhenyang *et al.*, 2017b], and conduct a comparative analysis against existing state-of-the-art trackers. The comparative outcomes are elucidated in Tab.1. It is worth noting that we only compared the results of single sampling during the testing phase. The effectiveness of multiple samplings has been demonstrated in our ablation study.

TNL2k. TNL2k is an extensive benchmark for large-scale natural language task tracking. Our proposed DMTrack achieves promising results on the TNL2k dataset, with scores of 57.7% on AUC and 59.9% on PRE.

LaSOT. The LaSOT dataset encompasses over 1,400 video sequences gathered from the internet, comprising a total of more than 3 million frames of visual images. Our proposed

Type	Method	Published	TNL2k		LaSOT		OTB-Lang		LaSOT$_{ext}$	
			AUC	PRE	AUC	PRE	AUC	PRE	AUC	PRE
Vision-Only	SiamFC[Bertinetto *et al.*, 2016]	ECCV2016	29.5	45.0	33.6	42.0	-	-	23.0	26.9
	SiamBAN[Chen *et al.*, 2022]	CVPR2020	41.0	48.5	51.4	59.8	-	-	-	-
	TransT[Chen *et al.*, 2021]	CVPR2021	50.7	57.1	64.9	73.8	-	-	-	-
	Mixformer[Cui *et al.*, 2022]	CVPR2022	-	-	69.2	78.7	-	-	-	-
	OSTrack[Ye *et al.*, 2022]	ECCV2022	54.3	-	69.1	78.7	-	-	47.4	53.3
	SeqTrack[Chen *et al.*, 2023]	CVPR2023	56.4	-	71.5	81.1	-	-	50.5	57.5
Vision&Vision-Language	VLT$_{TT}$[Guo *et al.*, 2022]	NeurIPS2022	53.1	53.3	67.3	72.1	76.4	93.1	48.4	55.9
	All-in-One[Chunhui *et al.*, 2023]	ACMMM2023	55.3	57.2	71.7	78.5	71.0	93.0	54.5	66.0
Vision-Language	TNLS[Zhenyang *et al.*, 2017a]	CVPR2017	-	-	-	-	55.0	72.0	-	-
	DAT[Xiao *et al.*, 2018]	Arxiv2018	-	-	27.0	30.0	65.0	89.0	-	-
	RTTNLD[Qi *et al.*, 2020]	WACV2020	25.0	27.0	35.0	35.0	61.0	79.0	-	-
	GTI[Yang *et al.*, 202]	TCSVT2021	-	-	47.8	47.6	58.1	73.2	-	-
	TNL2k-2[Xiao *et al.*, 2021]	CVPR2021	42.0	42.0	51.0	55.0	68.0	88.0	-	-
	SNLT[Qi *et al.*, 2021]	CVPR2021	27.6	41.9	54.0	57.6	66.6	80.4	-	-
	JointNLT[Li *et al.*, 2023]	CVPR2023	56.9	58.1	60.4	63.6	65.3	85.6	-	-
	DMTrack-256	**Ours**	**57.7**	**59.9**	**66.8**	**72.7**	**69.3**	**90.9**	**47.3**	**52.1**
	MMTrack-384[Zheng *et al.*, 2023]	TCSVT2023	58.6	59.4	70.0	75.7	70.5	91.8	49.4	55.3
	DMTrack-384	**Ours**	**61.6**	**65.4**	**70.1**	**76.5**	**71.2**	**92.3**	**51.1**	**58.3**

Table 1: Comparison on the TNL2k, LaSOT, OTB-Lang ,LaSOT$_{ext}$ test set with the state-of-the-art tracker. The vision-only type of method is evaluated by bounding box initialization, while the vision-language type of method is evaluated by a joint bounding box and natural language initialization. The best two results are highlighted in red and blue, respectively.

DMTrack performance on the LaSOT dataset is presented in Tab.1, achieving scores of 65.5% on AUC and 70.9% on PRE.

OTB-Lang. The OTB-Lang dataset comprises 99 video sequences, encompassing a diverse array of scenes and complexities. Our proposed DMTrack performance on the OTB-Lang dataset is presented in Tab.1, achieving scores of 68.5% on AUC and 89.7% on PRE.

LaSOT$_{ext}$. As the publicly released extension dataset of LaSOT, LaSOT$_{ext}$ comprises 150 challenging long-term videos from 15 object classes. The test results of our proposed DMTrack on the LaSOT$_{ext}$ dataset are presented in Tab.1, achieving a score of 45.3% on AUC and 51.1% on PRE.

4.3 Ablation Study

The effectiveness of the diffusion model. The diffusion model has been proven to efficiently capture semantic information from features in dense prediction tasks, providing pixel-level information for models. To validate the effectiveness of introducing the diffusion model, we compare it with a tracker that does not use the diffusion model. The comparison results, as shown in Tab.2, reveal that the tracker without the diffusion model exhibits a 2.1% decrease in AUC score, a 2.9% decrease in PRE score, and a 2% decrease in P score compared to the tracker with a single sampling of the diffusion model during the testing phase. In comparison, the tracker without the diffusion model shows a difference of 3.3% in AUC score, 4.6% in PRE score, and a 3.5% decrease in P$_{norm}$ score compared to the tracker with five samplings of the diffusion model during the testing phase. Our experimental results demonstrate the effectiveness of the proposed diffusion model.

The effectiveness of mask-driven. We attempted to remove the masking approach and solely use diffusion masking features for target localization, with the comparative results shown in Tab.2. We found that when using only diffusion masking features for target localization, as opposed to the mask-driven approach, the tracker experiences a 3.6% decrease in AUC score, a 4.2% decrease in PRE score, and a 3.5% decrease in P$_{norm}$ score. We analyze that the current data uses language descriptions generated based on the state or trend of the target in the first frame. However, over prolonged object motion, the language descriptions may become inaccurate, providing incorrect information. Therefore, our proposed tracker, utilizing a mask-driven approach, can more effectively utilize language descriptions for target tracking.

The effectiveness of language information. The key to visual-language tracking lies in the effective utilization of language features to enhance the tracking capability of traditional visual trackers in complex scenarios. To demonstrate that our tracker can effectively leverage language features, we trained a visual tracker without language features, and the results are presented in Tab.2. A pure visual tracker without language descriptions exhibits a 1.6% decrease in AUC score, a 1.1% decrease in PRE score, and a 1% decrease in P$_{norm}$ score compared to the visual-language tracker that utilizes language information. Through the comparison of experimental results, we validate that our proposed visual-language tracker can effectively leverage language features.

The effectiveness of candidate elimination module. To validate the effective elimination of ground truth background information by our proposed candidate elimination module and to enhance the sampling effectiveness of the diffusion model in the tracking task, we conducted experiments on a tracker without the candidate elimination module, and the experimental results are shown in Tab.2. The tracker without the candidate elimination module shows similar experimental performance in the initial sampling compared to using the candidate elimination module. However, after five samplings, the tracker without the candidate elimination module experienced a 1.2% decrease in the AUC score, a 2.1% decrease

#No.	Diffusion	Mask	Language	GT-CE	Train		Test		OTB-Lang		
					Single-Sample	Double-Sample	Single-Sample	Multiple-Sample	AUC	PRE	P_{norm}
1		✓	✓						67.2	88.0	82.5
2	✓		✓	✓		✓	✓		66.9	88.4	82.5
3	✓	✓		✓		✓	✓		68.7	89.8	83.5
4	✓	✓	✓			✓	✓		69.1	90.1	84.0
5	✓	✓	✓			✓		✓	69.3	90.5	84.3
6	✓	✓	✓	✓	✓		✓		69.0	90.3	84.2
7	✓	✓	✓	✓	✓			✓	67.8	89.6	83.2
8	✓	✓	✓	✓		✓	✓		69.3	90.9	84.5
9	✓	✓	✓	✓		✓		✓	70.5	92.6	86.0

Table 2: Ablation study of DMTrack on OTB-Lang. The best two results are highlighted in red and blue, respectively.

in the PRE score, and a 1.7% decrease in P_{norm} score compared to our tracker. The experiments demonstrate that our proposed candidate elimination module effectively improves the diffusion model's performance in the object-tracking task.

The impact of double-sample during the training phase. Traditional diffusion models employ a single-sampling method during the training phase, while we propose, for the first time, a double-sampling method during the training phase. The advantage of the double-sampling method lies in mitigating the over-reliance on the previous sampling results during multiple samplings in the testing phase, providing the diffusion model with more complex initial noisy images during the training phase, thus enhancing the robustness of the diffusion model in facing complex scenes. We compare a tracker trained using a single-sampling method with our tracker, as shown in Tab.2. During a five-sampling test phase, the tracker trained using a single-sampling method exhibits a 2.7% decrease in AUC score, a 3% decrease in PRE score, and a 2.8% decrease in P_{norm} score compared to our tracker. It is worth noting that the tracker trained using a single-sampling method experiences a 1.2% decrease in AUC score, a 0.7% decrease in PRE score, and a 1% decrease in P_{norm} score when conducting five samplings in the testing phase compared to a single-sampling. Through the comparison of experimental results, we validate the effectiveness of our proposed double-sampling training method during the training phase and the efficacy of our diffusion model.

5 Visualization and Limitations

5.1 Visualization

To validate the effectiveness of our proposed method, we visually present partial test results of DMTrack on the TNL2k dataset and compare them with existing advanced trackers, namely MMTrack, TNL2k-II, and VLT_{TT}. As illustrated in Fig.7, when confronted with complex scenes, our proposed diffusion mask-driven enables more accurate target localization compared to existing visual-language trackers. We also showcase the mask images obtained by integrating language information with visual cues for global target localization. Leveraging these mask images, our tracker demonstrates enhanced robustness in handling complex scenarios.

5.2 Limitations

Upon closer examination of experimental results, it becomes evident that the incorporation of language information results in a deterioration of performance in LaSOT compared to the visual-only model. The primary reason behind this lies in the nature of language descriptions in visual-language datasets, which are based on the state and trend of the target in the first frame. This description becomes inaccurate as the target moves, providing erroneous information that impacts the localization performance of visual-language trackers. We encourage scholars to delve further into methods for offering accurate language descriptions and more effective strategies for utilizing them in visual-language trackers.

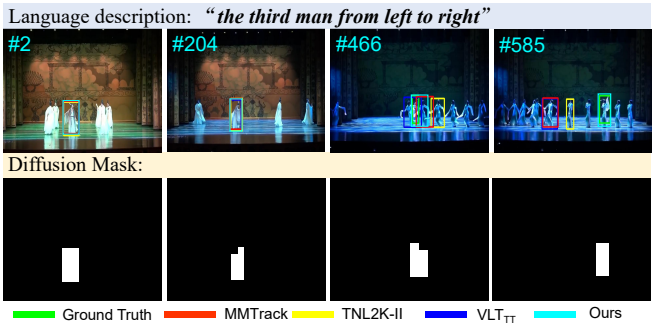

Figure 7: DMTrack is visually compared with three other VL trackers, namely MMTrack, VLT_{TT}, and TNL2k-II, on challenging sequences from the TNL2k benchmark.

6 Conclusion

In this work, we have observed that the initial language descriptions are often inaccurate in a highly time-varying video sequence and thus greatly deteriorate their tracking performance due to the low quality of extracted multi-modal features. In response to this challenge, we propose a visual-language tracker based on the diffusion model, which we refer to as the Diffusion Mask Visual-language Tracker. The core idea is to leverage the diffusion model to capture high-quality semantic information from low-quality multi-modal features and provide pixel-level target masks for visual features. Given the differences between dense prediction tasks and visual-language tracking tasks, we design a training approach involving double sampling, a feature filtering module, and the use of a ground truth candidate elimination module to adapt the diffusion model to visual-language tracking. Extensive experiments on multiple visual-language tracking benchmarks demonstrate the effectiveness of our approach.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.U23A20383), the Project of Guangxi Science and Technology (No.2024GXNSFGA010001 and 2022GXNSFDA035079), the Guangxi "Young Bagui Scholar" Teams for Innovation and Research Project, the Guangxi "Bagui Scholar" Teams for Innovation and Research Project, the Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing, the Guangxi Talent Highland Project of Big Data Intelligence and Application, and the Research Project of Guangxi Normal University (No.2022TD002).

References

[Alec *et al.*, 2021] Radford Alec, Kim JongWook, Hallacy Chris, Ramesh Aditya, Goh Gabriel, Agarwal Sandhini, Sastry Girish, Askell Amanda, Mishkin Pamela, Clark Jack, Krueger Gretchen, and Sutskever Ilya. Learning transferable visual models from natural language supervision. *arXiv,Cornell University*, 2021.

[Amit *et al.*, 2021] T. Amit, E. Nachmani, T. Shaharbany, and L. Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

[Bandara *et al.*, 2022] W. G. C. Bandara, N. G. Nair, and V. M. Patel. Ddpm-cd: Remote sensing change detection using denoising diffusion probabilistic models. *arXiv preprint arXiv:2206.11892*, 2022.

[Baranchuk *et al.*, 2021] D. Baranchuk, A. Voynov, I. Rubachev, V. Khrulkov, and A. Babenko. Label-efficient semantic segmentation with diffusion models. *International Conference on Learning Representations, ICLR*, 2021.

[Bertinetto *et al.*, 2016] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. *European Conference on Computer Vision,ECCV*, 2016.

[Chen *et al.*, 2021] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. *Conference on Computer Vision and Pattern Recognition, CVPR*, 2021.

[Chen *et al.*, 2022] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, Rongrong Ji, Zhenjun Tang, and Xianxian Li. Siamban: Target-aware tracking with siamese box adaptive network. *IEEE Transactions on Pattern Analysis and Machine Intelligence,TPAMI*, 2022.

[Chen *et al.*, 2023] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. *Conference on Computer Vision and Pattern Recognition, CVPR*, 2023.

[Chen *et al.*, 2024] Zhongxi Chen, Ke Sun, Xianming Lin, and Rongrong Ji. Camodiffusion: Camouflaged object detection via conditional diffusion models. *Association for the Advancement of Artificial Intelligence, AAAI*, 2024.

[Chunhui *et al.*, 2023] Zhang Chunhui, Sun Xin, Liu Li, Yang Yiqian, Liu Qiong, Zhou Xi, and Wang Yanfeng. All in one: Exploring unified vision-language tracking with multi-modal alignment. *ACM International Conference on Multimedia, ACMMM*, Jul 2023.

[Cui *et al.*, 2022] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. *Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

[Dhariwal and Nichol, 2021] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

[Dmitry *et al.*, 2022] Baranchuk Dmitry, Rubachev Ivan, Voynov Andrey, Khrulkov Valentin, and Babenko Artem. Label-efficient semantic segmentation with diffusion models. *International Conference on Learning Representations, ICLR*, 2022.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations,ICLR*, 2021.

[Fan *et al.*, 2018] 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. *Conference on Computer Vision and Pattern Recognition, CVPR*, 2018.

[Guo *et al.*, 2022] Mingzhe Guo, Zhipeng Zhang, Heng Fan, and Liping Jing. Divert more attention to vision-language tracking. In *NeurIPS*, 2022.

[Ho *et al.*, 2020] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[Jacob *et al.*, 2019] Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019.

[Junhua *et al.*, 2016] Mao Junhua, Huang Jonathan, Toshev Alexander, Camburu Oana, Yuille Alan, and Murphy Kevin. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[Lee *et al.*, 2022] S. Lee, H. Chung, J. Kim, and J. C. Ye. Progressive deblurring of diffusion models for coarse-to-fine image synthesis. *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

[Li *et al.*, 2022] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 2022.

[Li *et al.*, 2023] Zhou Li, Zhou Zikun, Mao Kaige, and He Zhenyu. Joint visual grounding and tracking with natural language specification. *Conference on Computer Vision and Pattern Recognition, CVPR*, Mar 2023.

[Qi *et al.*, 2020] Feng Qi, Ablavsky Vitaly, Bai Qinxun, Li Guorong, and Sclaroff Stan. Real-time visual object tracking with natural language description. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar 2020.

[Qi *et al.*, 2021] Feng Qi, Ablavsky Vitaly, Bai Qinxun, and Sclaroff Stan. Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021.

[Rahman *et al.*, 2023] A. Rahman, J. M., J. Valanarasu, I. Hacihaliloglu, and V. M. Patel. Ambiguous medical image segmentation using diffusion models. *the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2023.

[Song *et al.*, 2020] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[Tomer *et al.*, 2021] Amit Tomer, Nachmani Eliya, Shaharbany Tal, and Wolf Lior. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv:2112.00390*, 2021.

[Wang *et al.*, 2021] C. Wang, K. Yeo, X. Jin, A. C. Duarte, L. Klein, and B. Elmegreen. S3rp: Self-supervised super-resolution and prediction for advection-diffusion process. *Annual Conference on Neural Information Processing Systems*, 2021.

[Whang *et al.*, 2022] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar. Deblurring via stochastic refinement. *the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2022.

[Wolleb *et al.*, 2022] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin. Diffusion models for implicit image segmentation ensembles. *International Conference on Medical Imaging with Deep Learning*, 2022.

[Xiao *et al.*, 2018] Wang Xiao, Li Chenglong, Yang Rui, Zhang Tianzhu, Tang Jin, and Luo Bin. Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking. *Cornell University - arXiv,Cornell University - arXiv*, Nov 2018.

[Xiao *et al.*, 2021] Wang Xiao, Shu Xiujun, Zhang Zhipeng, Jiang Bo, Wang Yaowei, Tian Yonghong, and Wu Feng. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021.

[Yang *et al.*, 202] Z. Yang, T. Kumar, T. Chen, J. Su, and J. Luo. Grounding-tracking-integration. *IEEE Transactions on Circuits and Systems for Video Technology, TCSVT*, 202.

[Yang *et al.*, 2020] Zhengyuan Yang, Tushar Kumar, Tianlang Chen, Jingsong Su, and Jiebo Luo. Grounding-tracking-integration. *IEEE Transactions on Circuits and Systems for Video Technology, TCSVT*, 2020.

[Ye *et al.*, 2022] Botao Ye, Hong Chang, Bingpeng Ma, and Shiguang Shan. Joint feature learning and relation modeling for tracking: A one-stream framework. *European Conference on Computer Vision,ECCV*, 2022.

[Yihao *et al.*, 2022] Li Yihao, Yu Jun, Cai Zhongpeng, and Pan Yuwen. Cross-modal target retrieval for tracking by natural language. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun 2022.

[Zhang and Agrawala, 2023] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[Zheng *et al.*, 2023] Yaozong Zheng, Bineng Zhong, Qihua Liang, Guorong Li, Rongrong Ji, and Xianxian Li. Towards unified token learning for vision-language tracking. *IEEE Transactions on Circuits and Systems for Video Technology, TCSVT*, 2023.

[Zhenyang *et al.*, 2017a] Li Zhenyang, Tao Ran, Gavves Efstratios, Snoek Cees G. M., and Smeulders Arnold W. M. Tracking by natural language specification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[Zhenyang *et al.*, 2017b] Li Zhenyang, Tao Ran, Gavves Efstratios, Snoek Cees G. M., and Smeulders Arnold W. M. Tracking by natural language specification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.