Aware Distillation for Robust Vision-Language Tracking Under Linguistic Sparsity

Guangtong Zhang[1,2], Bineng Zhong[3], Shirui Yang[1], Yang Wang[1], Tian Bai[1]*

[1] The College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2] The College of Data Science and Artificial Intelligence, Jilin Engineering Normal University, Changchun, 130052, China
[3] The Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin, 541004, China
zhanggt25@mails.jlu.edu.cn, bnzhong@gxnu.edu.cn, {yangsr24, wyang24}@mails.jlu.edu.cn, baitian@jlu.edu.cn

Abstract

Vision-language object tracking overcomes the limitations of relying solely on visual features by leveraging language descriptions of objects to provide cross-modal semantic information, thereby enhancing model robustness in complex scenarios. However, most existing high-performance vision-language trackers are trained jointly on pure visual data and vision-language multimodal data. Due to the relative sparsity of language annotations in the data, the trackers tend to prioritize the localization role of visual features, diminishing the model's attention to language information. To mitigate this issue, we propose a novel vision-language tracker: Aware Distillation for Robust Vision-Language Tracking under Linguistic Sparsity (ADTrack). We introduce a knowledge distillation framework employing a knowledge-rich teacher model and a lightweight student model to establish modality correlations between vision and language, enabling efficient modeling between visual information and language descriptions. Specifically, our lightweight student module simultaneously distills language encoding capabilities from language models through teacher-guided learning on input language, while performing target-aware perception on template images using language descriptions to generate more effective template features for subsequent visual extraction. Furthermore, to ensure perceptual robustness in linguistically sparse scenarios, we simulate language-deficient conditions during training and employ contrastive learning to enhance model adaptability. Extensive experiments demonstrate that ADTrack reduces parameters by over 50% while achieving state-of-the-art (SOTA) performance and speed on vision-language tracking benchmarks, including LaSOT, LaSOText, TNL2K, OTB-Lang and MGIT.

Introduction

Vision-language object tracking(Yang et al. 2020; Wang et al. 2021) introduces language modality input beyond pure visual tracking(Ye et al. 2022). This description is typically manually generated to represent the target's state and motion trend in the first frame. However, this characteristic leads to two challenges for existing vision-language trackers: 1. Difficulty in language modality acquisition: The need for manual language description design complicates data collection and increases the difficulty of dataset construction. 2. Limited adaptability: Since descriptions are often based solely on the target's initial state and trend in the first frame, most language descriptions struggle to provide positive guidance when the target undergoes significant changes. With the continuous advancement of object tracking tasks, many excellent vision-language trackers have contributed methods to address these issues. Current mainstream solutions fall into two categories: 1. Joint Training (Guo et al. 2022; Zhang et al. 2023): Trackers are trained jointly on vision-language datasets and pure visual datasets, mitigating the scarcity of vision-language data caused by the difficulty of language annotation. 2. Leveraging Large Language Models (LLMs)(Sun et al. 2024; Li et al. 2025): Recent works utilize introduced large language models to generate or update target language descriptions, addressing the instability of language information caused by inaccurate descriptions.

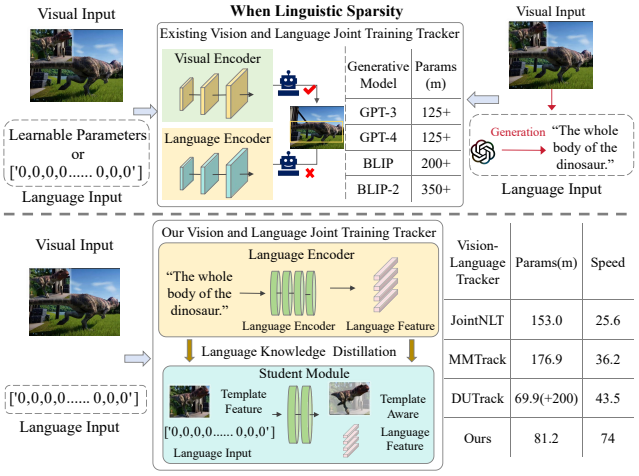

Figure 1: The comparison between our proposed ADTrack and traditional vision-language trackers. ADTrack alleviates the problem of trackers overly relying on visual features during joint training with fewer parameters, and provides the possibility for deployment extended to more scenarios.

*Tian Bai is the corresponding author.

While these methods effectively alleviate issues stemming from unstable language descriptions, they still lack more effective exploration is shown in the Fig.1. As illustrated: 1. Problem with Joint Training: An increasing number of vision-language trackers employ joint training on pure visual and vision-language data. Since pure visual datasets lack language descriptions, models often use an all-zero vector or a carefully designed learnable parameter to replace language input. Although this solves the input problem for the language encoder when annotations are absent, the generated " language features" typically lack substantive meaning. Consequently, the tracker becomes overly reliant on visual features during joint training, suppressing the significance of linguistic information. 2. Problem with LLM Integration: While LLMs like GPT demonstrate powerful advantages in many domains, their integration into trackers for generating target descriptions introduces new challenges. Although beneficial for providing annotations for pure visual data and alleviating description-target inconsistency during motion, LLMs are characterized by their massive parameter size and high computational cost. This significantly impacts tracker speed and hinders deployment in edge scenarios.

To address the issues of vision-language trackers overly focusing on visual information and failing to leverage the unique advantages of language in linguistically sparse scenarios, and to explore the possibility of extending the effective deployment of vision-language trackers to more scenarios, we propose Aware Distillation for Robust Vision-Language Tracking Under Linguistic Sparsity. We introduce knowledge distillation and contrastive learning, employing a carefully designed student model to learn from a pre-defined language teacher model. Unlike traditional knowledge distillation, our learning model effectively perceives language validity: 1. When language features are valid: They combine with language information to provide effective descriptions of the template image, enabling target awareness and reducing background interference. 2. When language features are invalid: The model leverages the template image's feature state to generate semantically meaningful language descriptions, mitigating the suppression of semantic information in linguistically sparse scenarios. Furthermore, our lightweight student model requires minimal complex design. This not only reduces more than 50% model parameters and accelerates tracking speed but also enables the potential deployment of future vision-language trackers in edge scenarios. The main contributions of our work are as follows:

- We propose a novel Aware Distillation framework for robust vision-language tracking under linguistically sparse scenarios. This approach mitigates the tendency of current vision-language trackers—trained jointly on pure visual unimodal data and vision-language multimodal data—to overemphasize the localization role of visual features, thereby enhancing the model's focus on language information.
- We introduce knowledge distillation and contrastive learning to design a lightweight student model. This model effectively perceives the validity of language information and leverages knowledge complementation between template features and language features.
- Extensive experiments demonstrate that our method, requiring minimal complex design, not only reduces tracker parameters but also achieves state-of-the-art results in both speed and performance. This provides the potential for deploying future vision-language trackers in edge scenarios.

Related Work

Vision-language Tracking

In recent years, the success of Vision Transformer (ViT (Dosovitskiy et al. 2021)) has validated Transformer's effectiveness in computer vision, while multi-modal applications—especially the fusion of language and vision—have gained growing attention (Alec et al. 2021). Early on, Li et al.(Li et al. 2017) proposed a natural language-based video object tracking method, introducing three models: one relying solely on language, one using vision after language-based initial location, and another combining both. Zhang et al. (Zhang et al. 2024a) suggested resolving the influence of inconsistent language information by exploiting the interaction between visual and language features. Shao et al. (Shao et al. 2024) obtained historical appearance features from past tracking results and produced precise language cues through the interplay of these historical features and language features. Sun et al. (Sun et al. 2024) put forward ChatTracker, which utilizes the extensive world knowledge in Multimodal Large Language Models (MLLMs) to generate high-quality language descriptions and boost tracking performance. Li et al. (Li et al. 2025) employed a large language model to create dynamic language descriptions of the target, aiming to optimize the initial language. Although these methods that introduce large models to generate language descriptions can alleviate the problem of visual data lacking language annotations or the issue of the model's attention bias caused by low-quality language annotations, one of their characteristics is that the models have a large number of parameters and high computational costs. This also leads to the fact that the introduction of large language generation models severely affects the speed of vision-language trackers and their deployment capabilities across multiple scenarios.

Knowledge Distillation

Knowledge Distillation (KD) enables lightweight networks (student networks) to be trained under the supervision of teacher models with more complex architectures and richer parameters. Through knowledge learning, it achieves excellent performance while reducing model costs. In multimodal knowledge distillation tasks, the concept of integrating multimodal information and improving learning effects through knowledge distillation methods has been successfully applied in multiple artificial intelligence fields (Li et al. 2020; Wang et al. 2023; Xu et al. 2021). Studies have shown that knowledge distillation in the multimodal domain can not only improve model performance but also enhance cross-modal relevance and generalization ability (Gou

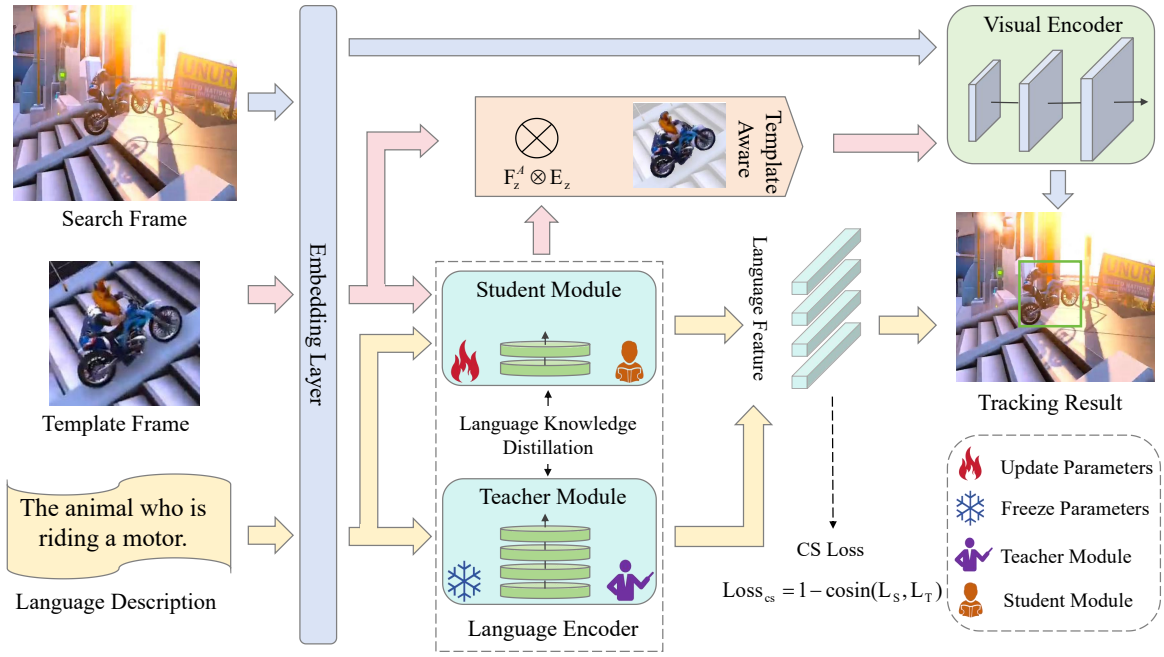

Figure 2: The overall structure of our proposed ADTrack. ADTrack is mainly composed of Visual Encoder, Language Encoder (comprising both Teacher Module and Student Module), and Multimodal Interaction Module.

et al. 2021; Huo et al. 2024). This technology has been applied and verified in various fields such as medical imaging (Wang et al. 2023) for filling modality gaps and action recognition (Radevski et al. 2023). The knowledge distillation method we introduce, through a carefully designed student model, can effectively perceive the validity of language. This method can combine multimodal information to establish cross-modal associations in different scenarios, provide target perception for visual information, and generate language descriptions with semantic information when language is sparse, thereby alleviating the suppression of semantic information by vision-language trackers in linguistically sparse scenarios.

Method

In this section, we will detail the specific components of the proposed ADTrack. The overall structure of our proposed ADTrack is shown in the Fig.2. First, we introduce the constituent modules of ADTrack, including: Visual Encoder, Language Encoder (comprising both Teacher and Student models) and Multimodal Interaction Module. Then, we describe the training procedure and corresponding loss functions. Finally, we outline the inference pipeline of ADTrack.

Visual Encoder

Most existing pure visual trackers adopt either a siamese or one-stream structure as the backbone encoder. Although the multimodal one-stream structure is simple and requires no complex design, the joint input of visual and language modalities typically demands greater computational resources due to the nature of attention mechanisms. Therefore, we select HiViT (Ghahremani et al. 2024) as ADTrack's visual encoder. Its inputs are the template image $z \in \mathbb{R}^{3\times H_z\times W_z}$ and the search image $x \in \mathbb{R}^{3\times H_x\times W_x}$. We first map both the template and search images into the same feature space via embedding layers.

$$\begin{aligned} E_z &= Embedding\left(z\right), \\ E_x &= Embedding\left(x\right). \end{aligned} \tag{1}$$

Crucially, we then linearly multiply the template-aware features F_z^A (obtained from the student model) with the original template features E_z to suppress background information in the template (details in Section *Language Encoder: Student Model*).

$$E_z' = F_z^A \times E_z. \tag{2}$$

Finally, the resulting template-aware features and the search features are jointly fed into Visual Encoder for visual feature extraction.

$$F_x = VisualEncoder\left(E_z', E_x\right). \tag{3}$$

Language Encoder

Teacher Module. Existing vision-language trackers primarily select language encoders from two categories: common language encoders (e.g., BERT(Devlin et al. 2019), RoBERTa(Liu et al. 2019)) or large language generation models (e.g., GPT(OpenAI 2023), BLIP(Li et al. 2022)). We aim for our proposed student module to maintain its lightweight nature, but its small parameter size makes it difficult to effectively learn from large language generation

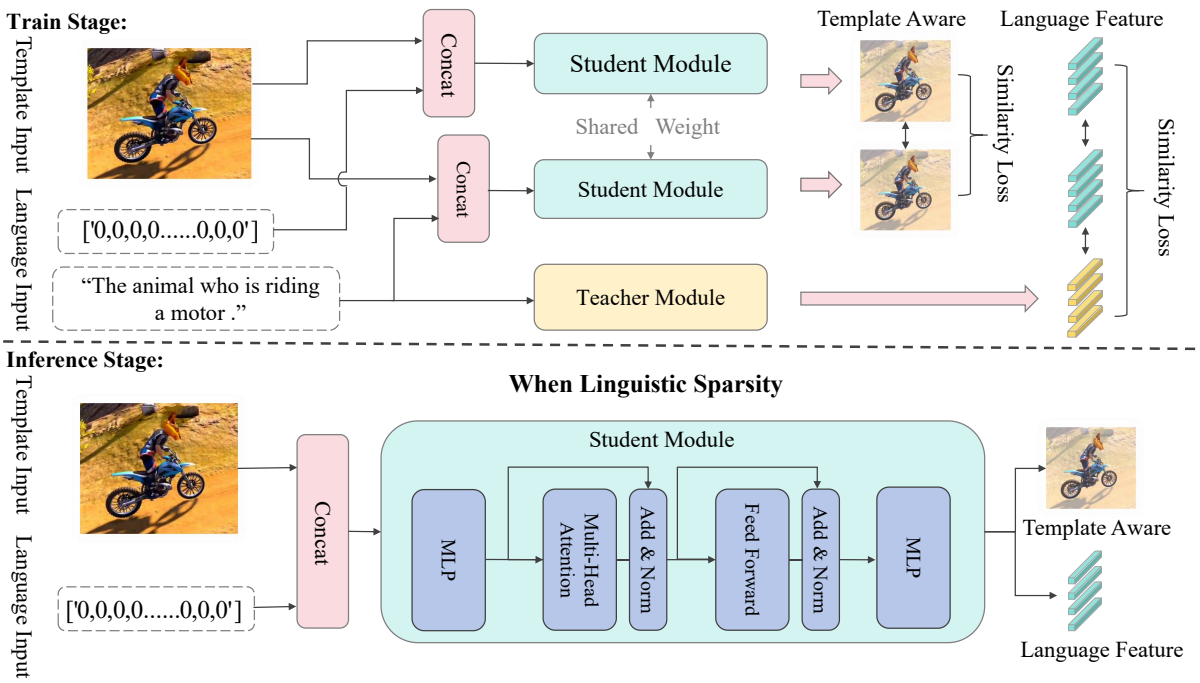

Figure 3: Data flow diagram of the training and inference phases of our proposed aware distillation method. In the training phase, we adopt a two-stage training approach to achieve effective deployment in linguistic sparsity scenarios; in the inference phase, ADTrack can realize rapid target localization.

models containing extensive knowledge. Therefore, we selected the common language encoder BERT as the teacher language encoder for our ADTrack. When presented with a natural language description as the query Q_l, our initial step is to tokenize the description and introduce CLS and SEP tokens. This leads to a token sequence represented as $L = \{\text{CLS}, \boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_N, \text{SEP}\}$, wherein N signifies the maximum permissible length of the language query. Following this, we feed the token sequence into our encoder to derive language token features denoted as $T_q \in \mathbb{R}^{C_q \times N}$, where $C_q = 768$ corresponds to the dimensionality of the output embedding.

$$F_t = TeacherModule\left(T_q\right). \tag{4}$$

Student Module. The architecture of our lightweight student module is illustrated in the Fig.3, primarily consisting of an " MLP + Global Attention + MLP" building block. Notably, our student module deviates from traditional knowledge distillation approaches where the input strictly matches that of the teacher module. To endow the student model with target-aware capability and effective language feature generation ability, the inputs to the student module include the template image z and the language description T_q (details in Section *Training Stage and Loss* and *Inference Stage*).

We first embed both the template image z and the language description T_q into a feature space using a shared embedding method. Then, the template embedding feature E_z and language embedding feature E_t are jointly fed into the student module, primarily focusing on feeding the generated template-aware features into the visual encoding layer. This process can be formulated as:

$$\begin{aligned} E_t, E_z &= Embedding\left(T_q, z\right), \\ F_t, F_z^A &= StudentModule\left(E_t, E_z\right). \end{aligned} \tag{5}$$

Multimodal Interaction Module

With the rapid development of vision-language multimodal tasks, diverse multimodal interaction methods have emerged across various domains. However, to demonstrate the simplicity and efficiency of our approach without necessitating excessive parameter design, the proposed ADTrack employs only cross-attention to achieve cross-modal interaction.

$$F_m = CrossAttention\left(F_x, F_t\right). \tag{6}$$

Training Stage and Loss

In this section, we mainly introduce the data flow and distillation process of the proposed ADTrack in the teacher module and student module. The proposed ADTrack adopts a two-stage training method. In the first stage, the training data of ADTrack consists entirely of vision-language data, meaning each video sequence is accompanied by a corresponding language description. The data flow of this process is shown in the Fig.3. The input to the teacher module is the language description of the sequence, which is used to generate effective language features.

$$F_t = TeacherModule\left(T_q\right). \tag{7}$$

The student module is divided into two steps. First, we use the same embedding method to embed the template image

z and language description T_q into the feature space. Then, we first send the template embedding feature E_t and a preset all-zero vector V_0 to the student module, focusing mainly on learning the generated language part features against the language features generated by the teacher module. After that, we send the template embedding feature E_z and language embedding feature E_t to the student module, with the focus on sending the generated template perception features F_z^A to the visual encoding layer. This process can be expressed as:

$$\begin{aligned} E_t, E_z &= Embedding\left(T_q, z\right), \\ F_t^0, F_z^s &= StudentModule\left(V_0, E_z\right), \\ F_t^s, F_z^A &= StudentModule\left(E_t, E_z\right). \end{aligned} \tag{8}$$

A cosine similarity loss is introduced into the loss function at this stage, the total loss can be expressed as:

$$\begin{aligned} L_{cosin}^t &= 1 - CosSimilar\left(F_t^0, F_t\right), \\ L_{cosin}^v &= 1 - CosSimilar\left(F_z^A, F_t^s\right), \\ Loss_{total} &= L_{cls} + L_{iou} + L_1 + L_{cosin}^t + L_{cosin}^v, \end{aligned} \tag{9}$$

where L_{cls} is the focal loss for classification, L_{iou} is the generalized IoU loss.

In the second stage, the training data of ADTrack includes pure visual data without language annotations and vision-language data with language annotations. In addition, in this stage, we freeze the parameters of the student module. For pure visual training data without language annotations, an all-zero vector is used to replace the language vector and sent to the student module together with the template features. The student module uses the template features as prompt information to generate effective language features. For vision-language data with language annotations, the language vector and template features are sent to the student module together. The student module uses the target description in the language information to suppress the background of the template image and achieve target perception. This step object tracking loss calculation is adopted in this stage, which can be expressed as:

$$Loss = L_{cls} + L_{iou} + L_1. \tag{10}$$

Inference Stage

During the inference phase of the model, we discard the teacher model and only use the student model. The template target perception features output by the student model are sent to the visual encoder to supress the background information of the template. Meanwhile, the language features it outputs are used as the language input of the multimodal interaction module, providing language semantic positioning for the tracker.

Experiments

Implementation Details

Model. In this selection, we introduce ADTrack experimental details. The feature embedding layers for template images and language respectively inherit the visual embedding layer of HiViT (Ghahremani et al. 2024) and the language embedding layer of Bert(Devlin et al. 2019). We have made public two variants of ADTrack based on different input sizes of visual images to meet the needs of different scenarios:

- ADTrack-256. The visual input to the 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 40, including a CLS and a SEP token.
- ADTrack-384. The visual input to the network is an image pair consisting of a template patch of size 192×192 and a search patch of size 384×384. For the language input, the max length of the language is set to 40, including a CLS and a SEP token.

Training Settings. The proposed ADTrack adopts a two-stage training process. In the first stage, we exclusively use vision-language tracking datasets with language annotations, including TNL2k(Wang et al. 2021), LaSOT(Fan et al. 2019), OTB-Lang(Li et al. 2017), and COCO-google(Mao et al. 2016). We employ AdamW to optimize the network parameters, with both the learning rate and weight decay set to 1×10^{-4}. During this stage, we train for 120 epochs, with a sample size of 60,000 images. In the second stage, the training data includes both vision-language tracking datasets with language annotations and pure visual tracking datasets without language annotations. Building upon the first stage, we augment the dataset with GOT-10k(Huang, Zhao, and Huang 2019) and TrackingNet(Muller et al. 2018). During this stage, we train for 80 epochs. In this stage, since pure visual tracking datasets do not contain language annotations, we use an all-zero vector as the language annotation. Moreover, during the training phase, the second stage loads the model parameters obtained from the first stage as pre-training and freezes the student model to ensure the validity of the model parameters.

State-of-the Art Comparisons

In this section, we compare our proposed ADTrack with existing excellent vision-language trackers and visual trackers on tracking benchmarks, with the results shown in the Tab.1.

TNL2k. The TNL2k(Wang et al. 2021) dataset is a large-scale, multi-modal dataset specifically designed for the field of vision-language tracking. The performance comparison of our proposed ADTrack on the test set of the TNL2k dataset is shown in Tab.1. Through comparisons with the performance of existing advanced trackers, our proposed ADTrack achieves significant performance advantages on the TNL2k benchmark.

LaSOT. The LaSOT(Fan et al. 2019) dataset is a large-scale, high-quality single-object tracking dataset designed for the field of visual tracking. The performance comparison of our proposed ADTrack on the test set of the LaSOT dataset is shown in Tab.1. Our proposed ADTrack has a slight gap compared to the best-performing model on the LaSOT benchmark.

OTB-Lang. The OTB-Lang(Li et al. 2017) dataset is an extended version of the traditional OTB dataset that adds

Type	Tracker	Source	TNL2K		LaSOT		OTB-Lang		LaSOT$_{ext}$		MGIT		FPS
			AUC	*Prec*	*AUC*	*Prec*	*AUC*	*Prec*	*AUC*	*Prec*	*SR*	*Prec*	
Vision-Only	SiamBAN(Chen et al. 2022)	CVPR2020	41.0	48.5	51.4	59.8	-	-	-	-	-	-	40
	TransT(Chen et al. 2021)	CVPR2021	50.7	57.1	64.9	73.8	-	-	-	-	44.7	53.9	50
	Mixformer(Cui et al. 2022)	CVPR2022	-	-	69.2	78.7	-	-	-	-	62.9	52.6	-
	OSTrack(Ye et al. 2022)	ECCV2022	54.3	-	69.1	78.7	-	-	47.4	53.3	58.3	47.6	105
	SeqTrack(Chen et al. 2023)	CVPR2023	56.4	-	71.5	81.1	-	-	50.5	57.5	-	-	15
	AQATrack-256(Xie et al. 2024)	CVPR2024	57.8	59.4	71.4	78.6	-	-	51.2	58.9	-	-	65
	DiffusionTrack-L256(Xie, Wang, and Ma 2024)	CVPR2024	56.4	57.3	70.8	76.7	-	-	-	-	-	-	45
	ARTrackV2-256(Bai et al. 2024)	CVPR2024	59.2	-	71.6	77.2	-	-	50.8	57.7	-	-	94
Vision-Language	TNLS-III(Li et al. 2017)	CVPR2017	-	-	-	-	55.0	72.0	-	-	-	-	-
	RTTNLD(Feng et al. 2020)	WACV2020	25.0	27.0	35.0	35.0	61.0	79.0	-	-	-	-	30
	GTI(Yang et al. 2020)	TCSVT2021	-	-	47.8	47.6	58.1	73.2	-	-	-	-	-
	SNLT(Feng et al. 2021)	CVPR2021	27.6	41.9	54.0	57.6	66.6	80.4	-	-	-	-	50
	TNL2K-2(Wang et al. 2021)	CVPR2021	41.7	42.0	51.0	55.0	68.0	88.0	-	-	-	-	25
	VLT$_{TT}$(Guo et al. 2022)	NeurIPS2022	53.1	53.3	67.3	72.1	76.4	93.1	48.4	55.9	47.4	32.4	35
	JointNLT(Zhou et al. 2023)	CVPR2023	56.9	58.1	60.4	63.6	65.3	85.6	-	-	60.3	43.3	39
	All-in-One(Zhang et al. 2023)	ACMMM2023	55.3	57.2	71.7	78.5	71.0	93.0	54.5	66.0	-	-	60
	DecoupleTNL(Ma and Wu 2023)	ICCV2023	56.7	56.0	71.2	75.3	73.8	94.8	-	-	-	-	32
	OSDT(Zhang et al. 2024a)	TCSVT2024	59.3	61.5	64.3	68.6	66.2	86.7	-	-	-	-	67
	DMTrack(Zhang et al. 2024b)	IJCAI2024	57.7	59.9	66.8	72.7	69.3	90.9	47.3	52.1	-	-	40
	UVLTrack-B(Ma et al. 2024)	AAAI2024	62.7	65.4	69.4	74.9	60.1	79.1	62.7	65.4	-	-	28
	QueryNLT(Shao et al. 2024)	CVPR2024	57.8	58.7	59.9	63.5	66.7	88.2	-	-	-	-	-
	DUTrack(Li et al. 2025)	CVPR2025	64.9	70.6	73.0	81.1	70.9	93.9	50.5	58.1	-	-	43
	ADTrack-256	**Ours**	**65.3**	**70.8**	**71.5**	**78.5**	**72.3**	**93.9**	**50.6**	**58.2**	**69.0**	**62.0**	**74**
	ADTrack-384	**Ours**	**66.1**	**71.5**	**72.6**	**80.3**	**73.4**	**95.5**	**51.2**	**59.1**	**70.0**	**64.9**	**41**

Table 1: Comparison of our method with state-of-the-art approaches on TNL2k(Wang et al. 2021), LaSOT(Fan et al. 2019), LaSOText(Fan et al. 2019), OTB-Lang(Li et al. 2017) and MGIT(Hu et al. 2023) datasets.

	Mixformer(Cui et al. 2022)	OSTrack(Ye et al. 2022)	AQATrack(Xie et al. 2024)	MIMTrack(Wang et al. 2025)	**ADTrack-256**	**ADTrack-384**
GOT10k	71.2	71.0	73.8	73.5	**74.8**	**76.1**

Table 2: Comparison with state-of-the-art methods on GOT-10k(Huang, Zhao, and Huang 2019) benchmarks in AO score.

language description information. The performance comparison of our proposed ADTrack on the test set of the OTB-Lang dataset is shown in Tab.1. Through comparisons with the performance of existing advanced trackers, our proposed ADTrack achieves significant performance advantages on the OTB-Lang benchmark.

LaSOText. The LaSOText(Fan et al. 2019) dataset is a high-quality benchmark dataset focusing on large-scale single object tracking. The performance comparison of our proposed ADTrack on the test set of the LaSOText dataset is shown in Tab.1. Through comparisons with the performance of existing advanced trackers, our proposed ADTrack achieves significant performance advantages on the LaSOText benchmark.

MGIT. The MGIT(Hu et al. 2023) is a multimodal video tracking evaluation benchmark for complex spatiotemporal causal relationships. The performance comparison of our proposed ADTrack on the test set of the MGIT dataset is shown in Tab.1. Through comparisons with the performance of existing advanced trackers, our proposed ADTrack achieves significant performance advantages on the MGIT benchmark.

GOT10k. GOT10k(Huang, Zhao, and Huang 2019) is a large-scale and highly diverse video dataset of object tracking algorithms. The performance comparison of our proposed ADTrack on the test set of the GOT10k dataset is shown in Tab.2. Although our proposed ADTrack is more committed to mitigating the model attention bias caused by data differences between pure visual and vision-language data, rather than directly generating high-quality language annotations like large language generation models, it still achieves certain advantages in pure visual benchmarks.

FPS. ADTrack employs knowledge distillation to significantly reduce model parameters and accelerate inference speed. A comparison of its performance, parameters, and FPS is shown in the Tab.3. Through comparison, ADTrack achieves leading performance with only half the parameters of existing vision-language trackers. In particular, ADTrack-256 reaches an FPS of 74, which is much faster than existing vision-language trackers, while ADTrack-384 achieves significant performance advantages at a similar speed.

Ablation Study

In this section focuses on designing ablation experiments and conducting analyses to verify the effectiveness of the knowledge distillation student module.

Study on Knowledge Distillation. Our proposed ADTrack introduces a novel lightweight knowledge distillation method. To verify the effectiveness of distillation learning, we designed an ablation experiment where knowledge compression of the teacher model is performed solely through distillation learning, with the results shown in the Tab.4. Although the performance of the tracker decreased by 0.4% after using distillation learning on the TNL2k benchmark, it reduced the parameters of the language encoder by over 90%, decreased the overall model parameters by more than

Tracker	Public	AUC(%)	Params(M)	Speed(FPS)
JointNLT	CVPR2023	56.9	153.0	25.6
MMTrack	TCSVT2023	58.6	176.9	36.2
DUTrack	CVPR2025	64.9	69.9(+200)*	43.5
ADTrack-256	**Ours**	**65.3**	**81**	**74**
ADTrack-384	**Ours**	**66.1**	**81**	**41**

Table 3: Comparison of performance, model parameters, and inference speed on TNL2k.(*DUTrack(Li et al. 2025) only discloses the number of parameters in the visual backbone part, and it uses Blip(Li et al. 2022) as its language annotation generation model. Therefore, we reasonably speculate that the overall parameters of its model exceed 200M.)

KD	Train Step	Generate Language	VA	AUC	Params	FPS
✗	One	✗	✗	63.4	165	51
✓	One	✗	✗	63.0	81	74
✓	Two	✗	✗	64.1	81	74
✓	Two	✓	✗	64.6	81	74
✓	Two	✓	✓	65.3	81	74

Table 4: Ablation study of ADTrack on TNL2k benchmark.

50%, and increased the speed by 23 FPS. This demonstrates the effectiveness of distillation learning in the field of vision-language tracking.

Study on Train Step and Generate Language Feature. Our proposed ADTrack adopts a two-stage training method. The first stage uses vision-language data with language annotations, and the second stage adds visual data without language annotations. To verify the performance changes after increasing the data volume and whether ADTrack's Language Feature Generate method can alleviate the modal attention bias caused by joint training, we design the following ablation experiments, with the results shown in the Tab.4. After adding visual data without language annotations, the model performance improved significantly: on the TNL2k benchmark, the AUC increased by 1% and the PR increased by 1.1%. However, when we used the Language Feature Generate method of the student model to mitigate the modal attention bias caused by joint training, the AUC increased by 1.6% and the PR increased by 2.5%. This confirms the effectiveness of our proposed method and also indicates the rationality of the problem of modal attention bias in the model caused by joint training.

Study on Visual Aware (VA). In the proposed ADTrack, the designed student model can not only generate language features in linguistically sparse scenarios but also utilize language features to achieve visual target perception when valid language annotations are available. To verify the effectiveness of the target perception method, we design an ablation experiment, with the results shown in the Tab.4. After adding target perception to our model, the AUC increased by 0.7% and the PR increased by 1.4% on the TNL2K benchmark, thus confirming the effectiveness of our method.

Figure 4: Visualisation comparison results of ADTrack with other vision-language trackers on challenging sequences from the TNL2k benchmark.

Visualization. The visualization comparison results of our tracker with four other trackers (i.e., JointNLT(Zhou et al. 2023), MMTrack(Zheng et al. 2023), UVLTrack(Ma et al. 2024), and DUTrack(Li et al. 2025)) on challenging sequences in the TNL2K benchmark are shown in the Fig.4. Our ADTrack demonstrates obvious advantages over existing excellent vision-language trackers in multiple complex scenarios, such as occlusion (Sequence 1), rapid changes (Sequence 2), similar object interference (Sequence 3), and target deformation (Sequence 4).

Conclusion

In this paper, we observe that vision-language trackers, due to being jointly trained on pure visual data without language annotations and vision-language multimodal data with language annotations, tend to overly focus on the localization role of visual features. To address this, we propose a novel vision-language tracker called Aware Distillation for Robust Vision-Language Tracking Under Linguistic Sparsity (ADTrack). By introducing knowledge distillation and contrastive learning methods, the carefully designed student model is endowed with target perception ability and effective language feature generation ability, enabling it to still have significant advantages in scenarios with Linguistic Sparsity. Extensive experiments show that our method not only significantly reduces model parameters but also achieves state-of-the-art performance and speed in vision-language tracking tasks.

Limitation. Our proposed ADTrack is still a certain distance to achieve effective deployment in constrained scenarios. In the future, we will consider adopting methods such as pruning or designing lightweight visual encoders and multimodal interaction module to realize the effective deployment of ADTrack in constrained scenarios (e.g., edge scenarios).

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants 62576149 and the Fundamental Research Funds for the Central University, JLU.

References

Alec, R.; JongWook, K.; Chris, H.; Aditya, R.; Gabriel, G.; Sandhini, A.; Girish, S.; Amanda, A.; Pamela, M.; Jack, C.; Gretchen, K.; and Ilya, S. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv,Cornell University*.

Bai, Y.; Zhao, Z.; Gong, Y.; and Wei, X. 2024. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19048–19057.

Chen, X.; Peng, H.; Wang, D.; Lu, H.; and Hu, H. 2023. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14572–14581.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8126–8135.

Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R.; Tang, Z.; and Li, X. 2022. SiamBAN: Target-aware tracking with Siamese box adaptive network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 5158–5173.

Cui, Y.; Jiang, C.; Wang, L.; and Wu, G. 2022. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13608–13618.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations,ICLR*.

Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; and Ling, H. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5374–5383.

Feng, Q.; Ablavsky, V.; Bai, Q.; Li, G.; and Sclaroff, S. 2020. Real-time visual object tracking with natural language description. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 700–709.

Feng, Q.; Ablavsky, V.; Bai, Q.; and Sclaroff, S. 2021. Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5851–5860.

Ghahremani, M.; Khateri, M.; Jian, B.; Wiestler, B.; Adeli, E.; and Wachinger, C. 2024. H-vit: A hierarchical vision transformer for deformable image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11513–11523.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*.

Guo, M.; Zhang, Z.; Fan, H.; and Jing, L. 2022. Divert more attention to vision-language tracking. *Advances in Neural Information Processing Systems*, 35: 4446–4460.

Hu, S.; Zhang, D.; Feng, X.; Li, X.; Zhao, X.; Huang, K.; et al. 2023. A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship. *Advances in Neural Information Processing Systems*, 36: 25007–25030.

Huang, L.; Zhao, X.; and Huang, K. 2019. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5): 1562–1577.

Huo, F.; Xu, W.; Guo, J.; Wang, H.; and Guo, S. 2024. C2kd: Bridging the modality gap for cross-modal knowledge distillation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *International Conference on Machine Learning, ICML*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. 2020. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*.

Li, X.; Zhong, B.; Liang, Q.; Mo, Z.; Nong, J.; and Song, S. 2025. Dynamic Updates for Language Adaptation in Visual-Language Tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19165–19174.

Li, Z.; Tao, R.; Gavves, E.; Snoek, C. G.; and Smeulders, A. W. 2017. Tracking by natural language specification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6495–6503.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ma, D.; and Wu, X. 2023. Tracking by natural language specification with long short-term context decoupling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14012–14021.

Ma, Y.; Tang, Y.; Yang, W.; Zhang, T.; Zhang, J.; and Kang, M. 2024. Unifying visual and vision-language tracking via contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4107–4116.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; and Ghanem, B. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, 300–317.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Radevski, G.; Luo, Z.; Zhu, Y.; Gool, L. V.; and Dai, D. 2023. Multimodal distillation for egocentric action recognition. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Shao, Y.; He, S.; Ye, Q.; Feng, Y.; Luo, W.; and Chen, J. 2024. Context-aware integration of language and visual references for natural language tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19208–19217.

Sun, Y.; Yu, F.; Chen, S.; Zhang, Y.; Huang, J.; Li, C.; Li, Y.; and Wang, C. 2024. ChatTracker: Enhancing Visual Tracking Performance via Chatting with Multimodal Large Language Model. *Advances in Neural Information Processing Systems, NeurIPS*.

Wang, H.; Ma, C.; Zhang, J.; Zhang, Y.; Avery, J.; Hull, L.; and Carneiro, G. 2023. Learnable crossmodal knowledge distillation for multi-modal learning with missing modality. *Medical Image Computing and Computer Assisted Intervention – MICCAI*.

Wang, X.; Nie, G.; Meng, J.; and Yan, Z. 2025. MIMTrack: In-Context Tracking via Masked Image Modeling. *Proceedings of the AAAI Conference on Artificial Intelligence, AAAI*.

Wang, X.; Shu, X.; Zhang, Z.; Jiang, B.; Wang, Y.; Tian, Y.; and Wu, F. 2021. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13763–13773.

Xie, F.; Wang, Z.; and Ma, C. 2024. Diffusiontrack: Point set diffusion model for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19124.

Xie, J.; Zhong, B.; Mo, Z.; Zhang, S.; Shi, L.; Song, S.; and Ji, R. 2024. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19300–19309.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Multimodal-cop: Self-supervised vision-language pre-training with auxiliary tasks. *arXiv preprint arXiv:2107.07773*.

Yang, Z.; Kumar, T.; Chen, T.; Su, J.; and Luo, J. 2020. Grounding-tracking-integration. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9): 3433–3443.

Ye, B.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2022. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European conference on computer vision*, 341–357. Springer.

Zhang, C.; Sun, X.; Yang, Y.; Liu, L.; Liu, Q.; Zhou, X.; and Wang, Y. 2023. All in one: Exploring unified vision-language tracking with multi-modal alignment. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5552–5561.

Zhang, G.; Zhong, B.; Liang, Q.; Mo, Z.; Li, N.; and Song, S. 2024a. One-stream stepwise decreasing for vision-language tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10): 9053–9063.

Zhang, G.; Zhong, B.; Liang, Q.; Mo, Z.; and Song, S. 2024b. Diffusion mask-driven visual-language tracking. In *Proc. 33rd Int. Joint Conf. Artif. Intell*, 1652–1660.

Zheng, Y.; Zhong, B.; Liang, Q.; Li, G.; Ji, R.; and Li, X. 2023. Toward unified token learning for vision-language tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4): 2125–2135.

Zhou, L.; Zhou, Z.; Mao, K.; and He, Z. 2023. Joint visual grounding and tracking with natural language specification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23151–23160.