big data and cognitive computing

MDPI

Article

Lightweight Multimodal Adapter for Visual Object Tracking

Vasyl Borsuk [1,*], Vitaliy Yakovyna [1,2] and Nataliya Shakhovska [1,*]

[1] Artificial Intelligence Department, Lviv Polytechnic National University, 12 Bandera Str., 79013 Lviv, Ukraine
[2] Faculty of Mathematics and Computer Science, University of Warmia and Mazury in Olsztyn, 2 Oczapowskiego Str., 10-719 Olsztyn, Poland
* Correspondence: vasyl.y.borsuk@lpnu.ua (V.B.); nataliya.b.shakhovska@lpnu.ua (N.S.)

Abstract

Visual object tracking is a fundamental computer vision task recently extended to multimodal settings, where natural language descriptions complement visual information. Existing multimodal trackers typically rely on large-scale transformer architectures that jointly train visual and textual encoders, resulting in hundreds of millions of trainable parameters and substantial computational overhead. We propose a lightweight **multimodal adapter** that integrates textual descriptions into a state-of-the-art visual-only framework with minimal overhead. The pretrained visual and text encoders are frozen, and only a small projection network is trained to align text embeddings with visual features. The adapter is modular, can be toggled at inference, and has negligible impact on speed. Extensive experiments demonstrate that textual cues improve tracking robustness and enable efficient multimodal integration with over 100× fewer trainable parameters than heavy multimodal trackers, allowing training and deployment on resource-limited devices.

Keywords: visual object tracking; multimodal learning; vision-language multimodality; transformers; parameter efficient fine-tuning; efficient deep learning

check for updates

Academic Editors: Hannan Azhar, Md Nazmul Huda and Hongying Meng

Received: 1 September 2025
Revised: 3 November 2025
Accepted: 13 November 2025
Published: 15 November 2025

Citation: Borsuk, V.; Yakovyna, V.; Shakhovska, N. Lightweight Multimodal Adapter for Visual Object Tracking. *Big Data Cogn. Comput.* **2025**, *9*, 292. https://doi.org/10.3390/bdcc9110292

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Visual object tracking (VOT) is a fundamental problem in computer vision with broad applications in video surveillance, human–computer interaction, augmented reality, and autonomous systems. In its classical formulation, a tracker is initialized with a visual exemplar (typically a bounding box in the first frame) and tasked with localizing the same target in subsequent video frames. While modern visual trackers have achieved remarkable accuracy and robustness, they remain limited by their reliance on purely visual initialization, which can be ambiguous in complex scenes with visually similar objects, occlusions, or significant appearance changes.

Over the past few years, visual-only trackers have undergone rapid advances, spanning both lightweight real-time models [1–3] and large-scale high-accuracy architectures [4–6]. However, due to the unstable reliability of RGB-only data in challenging scenarios, more and more studies are exploring the potential of multimodal data to address this issue.

Multimodal visual object tracking extends RGB-only paradigm by incorporating additional descriptions of the target [7–9]. Such cues can encode category, attributes, and context, guiding target localization even under visual ambiguity. Early work by [7] introduced the OTB-Lang dataset and demonstrated that language can assist recovery after tracking drift. Subsequent new datasets [8,9] enabled more robust benchmarks, while multimodal tracking models [10–12] leveraged cross-modal fusion to improve tracking accuracy. Despite these advances, most vision–language trackers rely on large-scale joint

training of visual and textual encoders, introducing significant computational cost and limiting their deployment on resource-constrained systems [13,14].

Parameter-efficient fine-tuning (PEFT) [15–17] offers a promising direction by freezing pretrained networks and learning only small adapter modules. ProTrack [18] introduced the prompting concept into the tracking algorithm by weighted addition of two modalities. ViPT [19] further improved this idea by learning a small number of prompt-learning parameters and SDSTrack [13] introduced adapter-based tuning. However, despite improvements, it still requires 14.79 M trainable parameters out of 107.80 M total parameters, which is far from lightweight.

A key difficulty in this area lies in integrating language information into high-performance visual trackers while avoiding substantial increases in model size or retraining overhead. Many existing methods achieve this by heavily altering the backbone architecture [11,13] or by redesigning fusion components [12]. These strategies typically keep both the vision and language modules unfrozen or introduce a large number of new trainable parameters, which reduces flexibility and complicates deployment in systems tailored for purely visual tracking. Additionally, most current trackers cannot be easily switched back to a vision-only configuration without retraining, limiting their usefulness in settings where language input is optional.

To address the aforementioned issues, this work introduces a small multimodal adapter that augments the vision-only tracker with an additional text input. The original visual tracking model and the language encoder remain entirely frozen while only a small projection network trains. This design yields negligible extra computational cost, and enables fast, adapter-only training. Importantly, the adapter can be seamlessly disabled, restoring the tracker to its original visual-only operation. In summary, the main contributions of this work are:

- We propose a compact adapter module that enables existing visual-only trackers to incorporate multimodal data. The adapter introduces a single linear projection layer that maps text embeddings into the feature space of the tracker's visual template. This design preserves the pretrained weights of both the visual backbone and the language encoder, allowing full multimodal integration without retraining large models.
- The proposed adapter adds only 0.461 million trainable parameters to the tracker making it suitable for real-time inference and even on-device fine-tuning on constrained hardware such as the NVIDIA Jetson Orin NX. This highlights the broader potential of our approach for deploying multimodal tracking systems in embedded and resource-limited environments.
- The modular structure of the proposed adapter allows it to be toggled on or off at inference: when disabled, the tracker operates in its original visual-only mode with no performance loss, ensuring backward compatibility with existing pipelines.

2. Related Work

Over the past few years, visual-only trackers have undergone rapid advances, spanning both lightweight real-time models and large-scale high-accuracy architectures. On the efficiency side, trackers such as HiT [1], FEAR [20], HCAT [2], and ref. [3] have demonstrated that carefully designed hybrid attention and convolutional components can deliver high tracking precision at low latency, making them suitable for deployment on resource-constrained devices. On the other end of the spectrum, large transformer-based trackers such as STARK [4], TransT [5], MixFormer [6], and its variants have shown that global self-attention mechanisms can significantly improve long-term robustness and accuracy, setting new state-of-the-art results on benchmarks like LaSOT [9] and GOT-10k [21]. Other notable developments include discriminative correlation filter–based hybrids [22,23], hierarchical

feature fusion approaches [2], and lightweight transformer designs that balance speed with accuracy [1,3]. Additionally, diffusion-based approaches [24] also demonstrated promising results. These innovations have pushed visual-only tracking to maturity, yet they remain fundamentally limited by their inability to incorporate high-level semantic cues that might help resolve visual ambiguities.

In recent years, multimodal visual object tracking has emerged as a compelling extension of this paradigm, allowing the target to be specified or supplemented with a natural language description. This vision–language tracking (VLT) approach introduces high-level semantic cues—such as object category, attributes, or contextual relationships—that may be difficult to infer from visual data alone. Early work by [7] first formulated a tracking-by-natural-language specification, creating the OTB-Lang benchmark and demonstrating that language can guide target localization and recovery after drift. Subsequent efforts introduced larger datasets such as TNL2K [8] and LaSOT [9], as well as stronger baselines like AdaSwitcher [8], which adaptively combines visual and language-guided search strategies.

Model architectures have evolved from CNN–LSTM fusion networks [7] to Siamese trackers with language-modulated templates (e.g., SNLT [10]), which achieve real-time performance without sacrificing accuracy, and to transformer-based joint vision–language models (e.g., JointNLT [11], MemVLT [12], CLDTracker [14]), which leverage deep cross-modal attention to align semantic and visual features. These methods have shown consistent gains on multimodal tracking benchmarks (e.g., LaSOT [9], GOT-10k [21], TNL2K [8]) and have inspired related work in referring video object segmentation, where models localize targets with pixel-level precision. However, state-of-the-art vision–language trackers often require substantial computational overhead and extensive training, making them less practical for lightweight or resource-constrained applications.

With the increasing scale of modern models [25,26], parameter-efficient fine-tuning (PEFT) has become an important strategy for adapting pre-trained networks. Initially proposed in natural language processing [15–17], PEFT has recently been widely explored in computer vision. Its central idea is to freeze the pre-trained backbone and train only a small set of additional parameters, making it particularly effective when available fine-tuning data is limited. Common approaches include adapter-based tuning, which inserts lightweight trainable adapters into frozen networks [15], and prompt tuning, which introduces learnable tokens to guide the model [17].

In multimodal visual object tracking, where annotated language–video data is scarce, full fine-tuning easily leads to overfitting and hinders the learning of generalizable cross-modal representations. PEFT offers a promising alternative, yet research in this area remains limited. ViPT [19] applied prompt tuning but suffered from modality gaps and an over-reliance on the primary modality. More recently, SDSTrack [13] attempted to mitigate overfitting and reduce the number of trainable parameters through parameter-efficient designs. However, despite improvements, it still requires 14.79 M trainable parameters out of 107.80 M total parameters, which is far from lightweight. Moreover, SDSTrack processes additional modalities by adapting RGB-based encoders, limiting its applications to only visual modalities like depth, thermal, etc. Both ViPT and SDSTrack are confined to visual-to-visual adaptation and cannot incorporate non-visual modalities, such as textual descriptions, which provide complementary semantic cues and enable a richer understanding of the target object beyond its visual appearance.

A central challenge in this field is how to incorporate language into high-performance visual trackers without significantly increasing model complexity or retraining cost. Existing approaches often modify the backbone [11,13] or fusion modules extensively [12], freezing neither the vision nor language components or introducing many additional trainable parameters, which limits flexibility and hampers deployment in systems optimized

for visual-only tracking. Furthermore, most current models cannot be trivially reverted to a visual-only mode without retraining, reducing their adaptability in scenarios where language input is optional.

3. Materials and Methods

To reduce the number of trainable parameters, mitigate the risk of overfitting [25,27,28], and simplify the training process, we propose a lightweight multimodal adapter that enables any pre-trained vision-only tracker to leverage multimodal information, such as natural language or categorical object descriptions. Unlike existing multimodal tracking algorithms [11,12,29] that jointly train large transformer-based vision-language architectures with hundreds of millions of parameters, our adapter introduces only a few additional parameters while keeping both the visual and textual encoders frozen.

The key idea is to embed auxiliary modality information (e.g., text) using a pretrained encoder, project it through a small learnable adapter into the same feature space as the visual template representation, and combine the two representations through simple element-wise addition. This design provides modularity and efficiency—the adapter can be seamlessly attached to any vision-only tracker to enable multimodal reasoning without altering the underlying architecture.

3.1. Vision-Only Tracking Formulation

In the standard visual object tracking setup, the goal is to localize a target object in each frame of a video sequence given an initial visual exemplar (template). Let the search frame be denoted as $x_{rgb} \in \mathbb{R}^{H_x \times W_x \times 3}$ and the template image as $z_{rgb} \in \mathbb{R}^{H_z \times W_z \times 3}$. The task of the vision-based tracker is to learn a function:

$$F_{rgb} : \{x_{rgb}, z_{rgb}\} \rightarrow B, \tag{1}$$

where B denotes the predicted bounding box of the target in the search frame.

Typically, the tracker F_{rgb} consists of two main components: a vision encoder E_{rgb} that extracts high-level feature representations, and a box prediction head H_{rgb} that regresses the target's location from the encoded features. Formally, the tracking process can be described as:

$$F_{rgb}(x_{rgb}, z_{rgb}) = H_{rgb}(E_{rgb}(x_{rgb}), E_{rgb}(z_{rgb})). \tag{2}$$

The encoder E_{rgb} produces compact feature representations for both the template and search images, which are then fused and processed by H_{rgb} to predict the target's bounding box in the search region.

3.2. Multimodal Extension via Adapter

We extend the above vision-only tracker into a multimodal tracker capable of integrating auxiliary information from an additional modality X (e.g., text or categorical data). Given template-side multimodal input z_X, the modality is first encoded by a pretrained encoder E_X, and the resulting embedding is projected into the visual feature space via a lightweight learnable adapter A_X. The projected multimodal feature is then combined with the visual template representation using element-wise addition. The resulting multimodal tracker can be expressed as:

$$F_{rgb,X}(x_{rgb}, z_{rgb}, z_X) = H_{rgb}\big(E_{rgb}(x_{rgb}),\ E_{rgb}(z_{rgb}) + A_X(E_X(z_X))\big). \tag{3}$$

Importantly , the only trainable component in this formulation is the adapter A_X, which contains a relatively small number of parameters (less than 0.5 M in our experi-

ments), while all other components—E_{rgb}, H_{rgb}, and E_X—remain frozen. This ensures low computational cost, fast convergence, and minimal risk of overfitting.

3.3. Integration with HiT Tracker

To demonstrate the effectiveness and generality of our approach, we instantiate the proposed multimodal adapter on top of the HiT hierarchical vision transformer framework [1]. In this setup, the vision encoder E_{rgb} corresponds to the LeViT encoder used in HiT, which extracts feature representations from both the template and search images. The box prediction head H_{rgb} is implemented as HiT's hierarchical transformer, consisting of a Bridge Module and a corner head, which together regress the target bounding box from the fused template and search features.

Our multimodal adapter integrates seamlessly into this pipeline by projecting the text embedding into the HiT feature space and adding it element-wise to the template representation. This modification introduces only 0.461 M additional parameters, keeping the model efficient and preserving its original inference speed. The resulting model enriches the tracker's understanding with semantic cues from textual descriptions while maintaining the original HiT functionality when the adapter is disabled. An overview of the proposed multimodal architecture training procedure is illustrated in Figure 1.

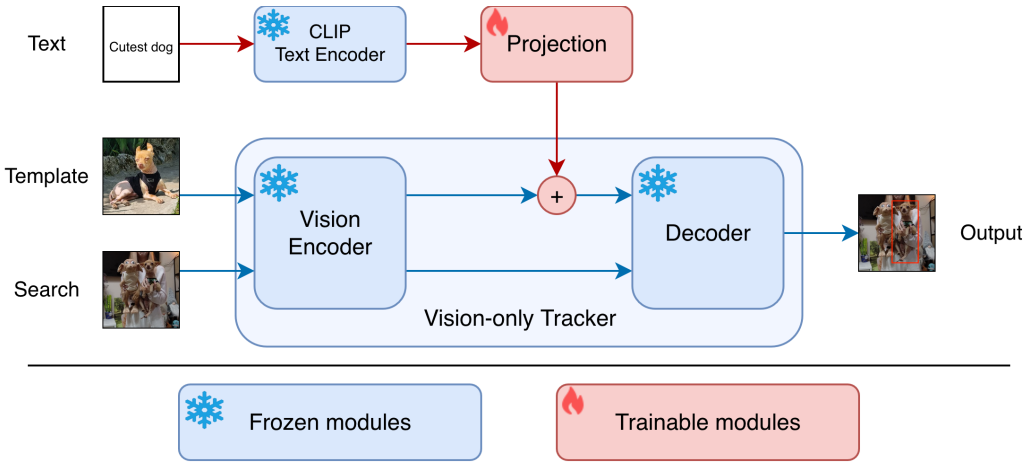


Figure 1. The overall training architecture of the proposed method. Blue components and arrows show the frozen parameters and computation flow of the original vision-only tracker. Red components and arrows mark the lightweight trainable adapter and the new multimodal operations it introduces. The plus sign (+) denotes element-wise addition.

3.4. Multimodal Adapter Details

We use the CLIP tiny text encoder [30] as E_X to extract fixed-dimensional embeddings from natural-language descriptions of target objects. During training, the CLIP encoder remains frozen to preserve its pretrained multimodal representations. The adapter A_X consists of two small linear projection layers followed by Batch Normalization [31] and a Hardswish activation function [32]. This simple projection is sufficient to align the text embedding with the template image features extracted by the HiT backbone (LeViT [33]), as shown in prior works [20]. Element-wise addition of the two feature vectors efficiently fuses visual and textual modalities without the need for complex cross-attention or fusion modules.

3.5. One-Hot Encoded Object Classes

In addition to more complex modalities like natural language, our framework supports a lightweight alternative modality based on one-hot encoded categorical labels. Each object class is represented by a sparse binary vector where only the entry corresponding

to the class is set to one. A small projection layer maps this vector into the same feature space as the template features $A_{cls}(E_{cls}(z_{cls})) \in \mathbb{R}^d$, where E_{cls} denotes the simple identity embedding of the one-hot vector. The projected class representation is added element-wise to the visual template features, following the same formulation as in the text-based adapter. Despite its extreme compactness—introducing only ∼0.043 M additional parameters compared to 0.461 M for the text adapter—the one-hot encoded version achieves competitive performance on the LaSOT benchmark (see Table 1), confirming the flexibility and efficiency of the proposed multimodal framework.

Table 1. Results of the proposed method using one-hot encoded object class information on the LaSOT benchmark.

Method	AUC	P_{Norm}	P	Trainable Params
HiT-Base	64.64	73.32	68.06	42.140 M
Ours (Text Adapter)	65.59	74.55	69.12	0.461 M
Ours (One-Hot Classes)	65.13	74.06	68.87	0.043 M

3.6. Training Procedure

All HiT visual object tracker and CLIP text encoder parameters are initialized from publicly released pretrained weights provided by the original authors of HiT [1] and CLIP [30]. Specifically, the HiT backbone is pretrained on large-scale visual tracking data as described in [1], while the CLIP text encoder is pretrained on 400 million image–text pairs. During our training, all pretrained parameters remain frozen (see Figure 1). The adapter's projection network parameters are the only trainable parameters and are initialized randomly. It results in only 0.461 million additional trainable parameters (entirely from the projection layer) and negligible additional computational overhead. The loss function is the same as in HiT and is equal to a combination of the ℓ_1 loss and the generalized intersection-over-union (GIoU) loss [34]:

$$\mathcal{L} = \lambda_G \mathcal{L}_{GIoU}(B, \hat{B}) + \lambda_l \mathcal{L}_l(B, \hat{B}). \quad (4)$$

where B denotes the ground-truth bounding box, and $\hat{B}$ denotes the predicted bounding box. The terms λ_G and λ_l are weighting coefficients that balance the contributions of the respective loss components and are equal to $\lambda_G = 2$ and $\lambda_l = 5$, respectively.

We used the AdamW optimizer [35] with a weight decay of 1×10^{-4}. The initial learning rate was set to 5×10^{-4}, with cosine decay over training. The batch size was 128, and the model was trained for 100 epochs, with each epoch containing approximately 60,000 sampled template–search pairs. This is substantially fewer epochs than the original HiT training schedule (1500 epochs), reflecting the efficiency of the adapter-only training. Training was performed on a single NVIDIA RTX A6000 GPU in mixed-precision mode. Additionally, such a small number of trainable parameters enables on-device training on resource-constrained hardware, the Nvidia Jetson Orin NX.

3.7. Inference

The proposed algorithm's pseudocode is presented in Algorithm 1. It is an end-to-end framework and does not require hyperparameter tuning. At the beginning of a video sequence, the template is initialized using the first frame along with its textual description encoded with the adapter. The search region is cropped for each subsequent frame based on the target's bounding box predicted in the previous frame. The template and search images are then input into our tracker, and the model directly outputs the bounding box of the target object.

When the multimodal adapter is enabled, the text description of the target object is embedded via CLIP [30], projected through the adapter, and combined with the template features. Importantly, the text embedding and projection are computed **only once during initialization**, exactly as template image features are extracted once and reused. This ensures no additional text processing overhead is incurred during subsequent frames. Because the adapter is modular, it can be **easily enabled or disabled at inference time**, preserving the original HiT model's visual-only performance when desired. If the adapter is disabled, the tracker reverts seamlessly to the original HiT visual-only pipeline. We do not use any additional post-processing methods such as window penalty or scale penalty, emphasizing that our framework is fully end-to-end and free of manually tuned hyperparameters.

Algorithm 1 Proposed algorithm.

Require: Frames $\{I_t\}_{t=1}^{T}$, initial box b_1, vision encoder f_θ, decoder d_θ
Require: Text description s_1, text adapter g_ϕ, UseAdapter $\in$ {True, False}
$z \leftarrow \text{CropAround}(I_1, b_1)$
$\phi_z \leftarrow f_\theta(z)$
if UseAdapter **then**
    $\phi_z \leftarrow \phi_z + g_\phi(s_1)$ ▷ applies multimodal adapter
end if
$\hat{b}_1 \leftarrow b_1$
for $t = 2$ to T **do**
    $x \leftarrow \text{CropAround}(I_t, \text{center}(\hat{b}_{t-1}))$
    $\phi_x \leftarrow f_\theta(x)$
    $\hat{b}_t \leftarrow d_\theta(\phi_z, \phi_x)$
end for
return $\{\hat{b}_t\}_{t=1}^{T}$

3.8. Datasets

The model was trained on the train splits of four publicly available datasets: TrackingNet [36], GOT-10k [21], LaSOT [9], and COCO2017 [37]. All datasets used in this study are publicly available, and their licensing terms permit research use. No proprietary data were used.

Video-based datasets (TrackingNet, GOT-10k, LaSOT): image pairs were sampled from random video sequences, with one frame used as the template and another as the search image.

Image-based dataset (COCO2017): a single image was randomly selected, and two views were generated via data augmentation to form the template and search pair.

Following [1], we extracted search regions and template patches by expanding the ground-truth bounding box by a factor of 4 for the search region and 2 for the template. Both were resized to fixed resolutions: 256 × 256 pixels for the search image and 128 × 128 pixels for the template image.

Data augmentation included random scaling, translation, and jittering, which were applied jointly to the template search pair to ensure spatial consistency.

4. Results

4.1. Evaluation on LaSOT Benchmark

We evaluated the proposed multimodal adapter for the HiT tracker on the LaSOT benchmark [9]. LaSOT is a large-scale long-term dataset consisting of 1400 video sequences, with 1120 for training and 280 for testing. Performance was measured using the standard tracking metrics: AUC (Area Under the Curve), precision (P), and normalized precision (P_{Norm}).

The AUC metric is derived from the success plot, which depicts the proportion of frames whose Intersection over Union (IoU) between the predicted and ground-truth

bounding boxes exceeds varying thresholds. A higher AUC value indicates better overall overlap accuracy across different IoU thresholds. Precision measures the percentage of frames where the center location error between prediction and ground truth is within a fixed threshold of 20 pixels, thus reflecting the spatial localization accuracy. Normalized precision provides a scale-independent version of precision by normalizing the distance with respect to the target size, offering a fairer comparison across sequences with different resolutions. All metrics are computed as defined in [9].

Table 2 reports the comparison between our method and state-of-the-art trackers. Our multimodal adapter improves the HiT baseline from 64.6 to 65.6 AUC, from 68.1 to 69.1 precision, and from 73.3 to 74.6 normalized precision, while not affecting the inference speed and having 90 times fewer trainable parameters than when training the whole network. This demonstrates that introducing text-based features via the adapter provides consistent improvements across all LaSOT metrics while keeping the model lightweight.

In addition, the proposed adapter improves HiT-Base and enables it to outperform large transformer networks like TransT [5]. While the original HiT-Base lagged behind TransT by 0.3 AUC, 0.5 normalized precision, and 0.9 precision, the adapter makes it better than TransT by 1.0 AUC, 0.8 normalized precision, and 0.1 precision. Moreover, our method outperforms lightweight trackers such as FEAR [20] and LightTrack [3] by large margins (over +12% AUC), while maintaining real-time inference speed and having 3–4 times fewer trainable parameters.

Compared with large multimodal approaches such as MemVLT [12] and MMTrack [38], our method achieves competitive accuracy with less than 1% of their trainable parameters, highlighting its efficiency.

Table 2. Performance comparison on the LaSOT benchmark. The proposed multimodal adapter enhances the HiT tracker with minimal additional parameters.

Method	AUC	P_{Norm}	P	FPS	Params	Trainable Params
Ours	65.6	74.6	69.1	173	42.67 M	0.46 M
HiT-Base [1]	64.6	73.3	68.1	175	42.14 M	42.14 M
TransT [5]	64.9	73.8	69.0	63	58 M	58 M
MixFormer-L [6]	70.1	79.9	76.3	18	183 M	183 M
MemVLT [12]	72.9	85.7	80.5	24	175 M	175 M
MMTrack [38]	70.0	82.3	75.7	29	177 M	177 M
JointNLT [11]	60.4	69.4	63.6	31	153 M	153 M
FEAR [20]	53.5	-	-	114	1.37 M	1.37 M
LightTrack [3]	53.8	-	-	128	1.97 M	1.97 M
HCAT [2]	59.3	68.7	61.0	195	-	-

4.2. Model Efficiency

A key advantage of the proposed adapter is its efficiency. From the Table 2, the adapter introduces only 0.461 M trainable parameters (about 1% of HiT's 42 M parameters). Unlike existing multimodal trackers that retrain hundreds of millions of parameters, our training requires only a small projection layer, making optimization much faster and less resource-intensive.

In terms of speed, the adapter-equipped model runs at 173 FPS on an NVIDIA GeForce RTX 2080 GPU, nearly identical to the HiT baseline (175 FPS). Thus, the proposed design integrates text features without affecting real-time performance, which is crucial for practical applications.

4.3. Ablation Study

We additionally provide an ablation study that proves the efficiency of the proposed framework. The Table 3 demonstrates that the proposed training procedure with frozen vision tracker parameters and trainable projection layers converges faster and produces more accurate results than training all model parameters after the same number of epochs. It reaches 65.59 AUC compared to only 65.06 while having more than 90 times fewer trainable parameters.

When training the entire HiT tracker together with the adapter, the model achieves slightly higher accuracy, but only after 4 times more training epochs and with 90× more trainable parameters. Such model training yields diminishing returns relative to the computational overhead.

In addition, we experiment with an alternative multimodal features combination strategy inspired by prompt learning approaches, where extra modality features are concatenated to the visual template features rather than added. Interestingly, this approach produced almost identical accuracy to our additive fusion design, yet required significantly more computation due to the increased sequence length in the transformer box prediction head. Specifically, it increases overall MACs (multiply-accumulate operations) of the algorithm from 4.35 G to 4.52 G. Moreover, concatenation-based integration complicates positional encoding alignment between modalities, making the training pipeline less stable and model-dependent. For these reasons, we adopt the additive fusion scheme as the preferred design choice throughout this work.

Table 3. Ablation study comparing the proposed method with frozen HiT visual tracker and trainable projection layers to variants with full fine-tuning and prompt learning concatenation of extra modality. Results are reported on the LaSOT benchmark.

Method	AUC	P_{Norm}	P	Trainable Params
HiT-Base	64.64	73.32	68.06	42.140 M
Ours	65.59	74.55	69.12	0.461 M
Prompt learning concat	65.61	74.52	69.08	0.461 M
All trainable	65.06	73.69	68.69	42.601 M
All trainable ×2 epochs	65.31	73.95	68.84	42.601 M
All trainable ×3 epochs	65.54	74.43	69.02	42.601 M
All trainable ×4 epochs	65.63	74.57	69.15	42.601 M

4.4. One-Hot Encoded Object Classes

We further experiment with a simplified setting where natural language descriptions are replaced by one-hot encoded object class information. This setup reduces the framework to an even more lightweight form, requiring only a negligible number of trainable parameters while still providing competitive accuracy. As shown in Table 1, the one-hot encoded framework achieves 65.13 AUC, 74.06 P_{Norm}, and 68.87 P, which is only slightly below the performance of the full text adapter while being more efficient in terms of additional parameters. This indicates that even coarse categorical cues can improve the baseline HiT tracker and that our framework is flexible enough to incorporate both rich textual and simple categorical object information.

4.5. Adapter in Larger Networks

To further evaluate the generality of the proposed framework, we applied the multimodal adapter to a larger and more powerful transformer-based tracker, MixFormer-L [6]. Table 4 summarizes the results on the LaSOT benchmark. The baseline MixFormer-L achieves 70.1 AUC, 79.9 normalized precision, and 76.3 precision. When equipped with our lightweight text adapter, the performance improves to 71.9 AUC, 83.6 normalized precision,

and 79.1 precision. Notably, this gain is achieved with only 0.659 M additional trainable parameters, which is negligible compared to the 183 M parameters of the backbone network. Moreover, the adapter-enhanced MixFormer-L surpasses the performance of MMTrack [38], which reports 82.3 normalized precision, while requiring 177 M trainable parameters.

These results demonstrate that the proposed adapter is not only effective for lightweight trackers like HiT, but can also enhance the performance of large-scale state-of-the-art architectures such as MixFormer-L, confirming the flexibility and scalability of our approach.

Table 4. Results of the proposed method when applied to a larger network (MixFormer-L) on the LaSOT benchmark.

Method	AUC	P_{Norm}	P	Trainable Params
MixFormer-L	70.1	79.9	76.3	183 M
MixFormer-L with Text Adapter	71.9	83.6	79.1	0.659 M

4.6. Stability Analysis via K-Fold Cross-Validation

To assess the robustness and stability of the proposed multimodal adapter, we performed a k-fold experiment with $k = 5$. The dataset was randomly divided into five folds while preserving the overall distribution of object classes and sequence durations. In each iteration, four folds were used for training, and the model was tested on the LaSOT benchmark.

Figure 2 illustrates the variability of AUC, normalized precision (P_{Norm}), and precision (P) across the five folds. The results exhibit minimal variance, and the corresponding standard deviations are below 0.04 for AUC, 0.09 for P_{Norm}, and 0.16 for P, confirming that the observed performance improvements are statistically stable and not dependent on a specific training split.

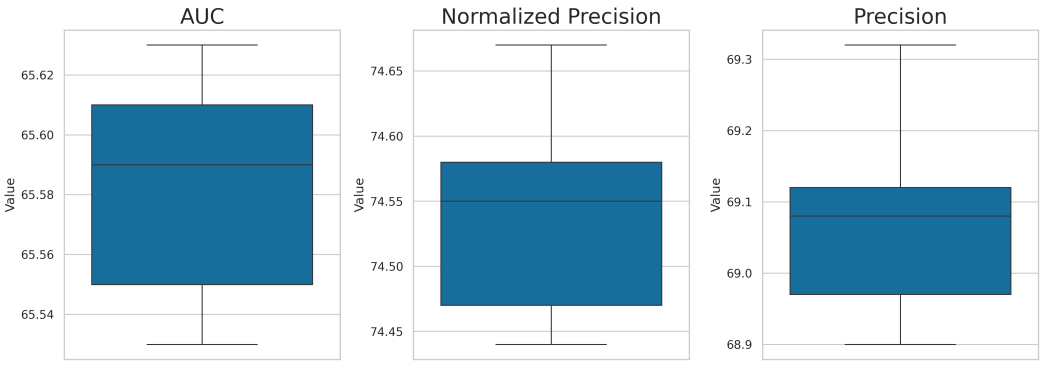


Figure 2. Box–whisker plots of AUC, normalized precision, and precision on the LaSOT benchmark for the proposed method over five folds. The narrow interquartile ranges indicate consistent performance across different data partitions.

Overall, the 5-fold experiment validates the generalization capability of the adapter-equipped tracker. The consistent median values across folds demonstrate that the proposed approach yields reproducible improvements, further reinforcing its reliability and robustness on large-scale tracking benchmarks.

4.7. Visualization of Tracking Results

To complement the quantitative results, we present a qualitative visualization that illustrates the effect of integrating the additional textual modality into the visual tracking framework. Figure 3 compares the output scoremaps of the baseline HiT tracker and the proposed multimodal adapter across several challenging scenarios.

As shown in Figure 3, the baseline HiT tracker often becomes confused when multiple visually similar objects appear in the same scene or template frame is ambiguous. For example, when several objects share similar appearance and motion patterns, the visual-only model may drift toward the wrong instance after partial occlusion or abrupt motion. Additionally, vision-only tracker fails to detect target object when template image is blurred or low quality during the initialization. In contrast, the proposed multimodal adapter utilizes the additional textual cue to disambiguate between such instances by conditioning the template representation on the target's semantic description. This additional modality effectively guides the model to focus on the correct object, thereby maintaining consistent localization throughout the sequence.

These qualitative results highlight how multimodal integration not only improves numerical performance but also leads to more interpretable and reliable tracking behavior. By reducing ambiguity in complex visual scenes, the adapter enables the tracker to remain robust even in cases where purely visual features are insufficient for discriminative target identification.



Figure 3. Qualitativecomparison of scoremap outputs on the search image between the baseline HiT tracker and the proposed multimodal adapter. The adapter leverages textual descriptions of the target to resolve ambiguities among visually similar objects or ambiguous template frames. Red rectangles denote the initialization bounding boxes.

4.8. Experimental Conclusions

The experimental findings can be summarized as follows:

- Adding a lightweight multimodal adapter consistently improves tracking accuracy across all LaSOT metrics;
- The adapter introduces only 0.461 M additional trainable parameters (over 90× fewer than training a full multimodal network end-to-end), enabling efficient optimization;
- A simplified variant using one-hot encoded object class information further demonstrates the flexibility of the framework. Despite requiring fewer trainable parameters, it achieves comparable accuracy on the LaSOT benchmark, confirming that even coarse categorical cues can enhance the visual-only tracker;
- The small number of trainable parameters makes the approach suitable for on-device training on resource-constrained platforms such as the Nvidia Jetson Orin NX;
- Real-time tracking performance is preserved at 173 FPS, making the approach suitable for practical deployment.

- When applied to larger architectures such as MixFormer-L, the adapter achieves substantial accuracy improvements and enables the tracker to outperform strong multimodal baselines like MMTrack while requiring orders of magnitude fewer trainable parameters.

Overall, these results validate the effectiveness of the proposed adapter as a simple yet powerful way to integrate textual descriptions into existing vision-only trackers. It improves robustness and accuracy while retaining the speed and efficiency of the baseline model.

5. Discussion

The results presented in Section 4 demonstrate that our proposed multimodal adapter provides consistent improvements over the baseline HiT tracker while preserving real-time performance and introducing only a negligible number of additional trainable parameters. In particular, the adapter improves the AUC on the LaSOT benchmark from 64.6 to 65.59, along with gains in both precision and normalized precision. These findings suggest that even a lightweight integration of textual information can yield measurable benefits in visual object tracking tasks. Moreover, due to its compact design, the proposed framework can be trained on resource-constrained edge devices such as Nvidia Jetson Orin NX.

5.1. Comparison with Prior Work

Previous multimodal tracking approaches, such as MTTR [29], MemVLT [12], and MMTrack [38], rely on large-scale transformer architectures that jointly train vision and language encoders, or introduce complex memory modules. While these methods achieve strong results, they require training hundreds of millions of parameters and significant computational resources, and often sacrifice inference speed. By contrast, our adapter design offers a more efficient alternative, requiring less than 1% of the parameters while still producing competitive accuracy.

Furthermore, unlike ViPT [19] and SDSTrack [13], which are limited to visual-only modalities such as RGB, depth, or thermal inputs, our method seamlessly incorporates textual information as an additional modality. This enables our tracker to benefit from high-level semantic cues provided by natural language, improving discrimination between visually similar targets and enhancing robustness under challenging conditions.

Importantly, even though the baseline HiT tracker operates with fewer parameters than large-scale vision-only models such as TransT [5], it initially underperforms them in purely visual settings. However, when augmented with our proposed multimodal adapter, HiT not only closes this gap but surpasses larger vision-only trackers—achieving superior accuracy while incurring negligible computational overhead. This demonstrates that introducing textual information through a parameter-efficient mechanism can deliver substantial performance gains without compromising efficiency.

5.2. Interpretation of Findings

The ablation study further emphasizes the efficiency of the adapter framework. Training only the small projection layers not only reduces optimization costs but also converges faster and achieves better accuracy compared to retraining the full HiT tracker for the same number of epochs. This suggests that when combined with the pre-trained multimodal representations such as CLIP [30] or one-hot encoded object class information, the frozen vision-only tracker already contains sufficient complementary information. The adapter's role is primarily to effectively align these feature spaces, which can be achieved with minimal trainable capacity.

These results support the hypothesis that text guidance provides useful high-level semantic information that purely visual features cannot capture. For instance, descriptions like "the red car in the left lane" or "the person wearing a blue jacket" encode discriminative cues that help the tracker distinguish between visually similar targets. Our approach demonstrates that such cues can be injected into existing trackers without the need for costly architectural redesigns.

5.3. Implications and Broader Context

The ability to seamlessly enable or disable the adapter at inference time is significant from a practical standpoint. It allows practitioners to preserve the original visual-only performance of HiT when no textual input is available, while leveraging multimodal cues when they are provided. This modularity makes the approach attractive for real-world applications such as human–robot interaction, video surveillance, or autonomous navigation, where user input in natural language can enhance robustness and interpretability.

The broader implication is that multimodal integration in tracking does not necessarily require large-scale end-to-end training. Instead, lightweight adapter mechanisms may provide a more sustainable and resource-efficient path forward, particularly for deployment on resource-constrained hardware such as mobile devices, drones, or embedded systems.

5.4. Future Research Directions

Although promising, our work also opens several avenues for further research:

- **Richer multimodal cues**: Beyond textual descriptions, future adapters could incorporate audio signals, depth, or other sensory modalities to improve robustness further. Moreover, multiple adapters can be used simultaneously.
- **Dynamic adapter mechanisms**: Instead of a fixed projection, adapters that dynamically weight text features based on the video context may yield more substantial improvements.
- **Dataset diversity**: Our experiments were conducted primarily on LaSOT; extending evaluations to more diverse multimodal benchmarks could reveal additional insights into generalization.
- **User studies**: Investigating how human users interact with language-guided trackers in real-world tasks could validate the practical utility of the approach.

6. Conclusions

This work introduced a lightweight multimodal adapter for vision-only object tracker, enabling the seamless integration of textual descriptions into a state-of-the-art visual-only framework. The adapter adds only 0.461 M trainable parameters and preserves real-time inference speed, while consistently improving accuracy across standard tracking metrics. In addition, the proposed design is suitable for inference and training on resource-constrained edge-devices.

Compared to existing multimodal tracking methods that rely on large-scale architectures and heavy retraining, our approach offers a significantly more efficient alternative. By freezing the vision and text encoders and training only a small projection layer, we demonstrate that multimodal capabilities can be achieved without substantial computational or parameter overhead.

The findings highlight the value of language as a complementary modality in object tracking, showing that even simple natural-language cues can enhance robustness in challenging scenarios. More broadly, this work suggests that lightweight adapter mechanisms provide a promising pathway for bringing multimodal intelligence to practical, resource-constrained applications.

Author Contributions: Conceptualization, V.B.; methodology, V.B.; software, V.B.; validation, V.Y. and N.S.; formal analysis, V.Y. and N.S.; investigation, V.B. and V.Y.; resources, V.B. and N.S.; data curation, V.B.; writing—original draft preparation, V.B. and V.Y.; writing—review and editing, V.B. and V.Y.; visualization, V.B.; supervision, V.Y. and N.S.; project administration, V.Y.; funding acquisition, N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by British academy fellowship grant number RaR\100727

Data Availability Statement: The raw data used in this study are available in [9,21,36,37].

Acknowledgments: During the preparation of this manuscript, the author used ChatGPT (GPT-4o, OpenAI) for the purposes of searching for relevant literature related to the topic and validating the clarity and consistency of the text. The authors have reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kang, B.; Chen, X.; Wang, D.; Peng, H.; Lu, H. Exploring lightweight hierarchical vision transformers for efficient visual tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 9612–9621.
2. Chen, X.; Kang, B.; Wang, D.; Li, D.; Lu, H. Efficient visual tracking via hierarchical cross-attention transformer. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 461–477.
3. Yan, B.; Peng, H.; Wu, K.; Wang, D.; Fu, J.; Lu, H. Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15180–15189.
4. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning spatio-temporal transformer for visual tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10448–10457.
5. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 8126–8135.
6. Cui, Y.; Jiang, C.; Wang, L.; Wu, G. Mixformer: End-to-end tracking with iterative mixed attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13608–13618.
7. Li, Z.; Tao, R.; Gavves, E.; Snoek, C.G.; Smeulders, A.W. Tracking by natural language specification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6495–6503.
8. Wang, X.; Shu, X.; Zhang, Z.; Jiang, B.; Wang, Y.; Tian, Y.; Wu, F. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13763–13773.
9. Fan, H.; Bai, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Harshit; Huang, M.; Liu, J.; et al. Lasot: A high-quality large-scale single object tracking benchmark. *Int. J. Comput. Vis.* **2021**, *129*, 439–461. [CrossRef]
10. Feng, Q.; Ablavsky, V.; Bai, Q.; Sclaroff, S. 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, Nashville, TN, USA, 19–25 June 2021; pp. 5851–5860.
11. Zhou, L.; Zhou, Z.; Mao, K.; He, Z. Joint visual grounding and tracking with natural language specification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 23151–23160.
12. Feng, X.; Li, X.; Hu, S.; Zhang, D.; Zhang, J.; Chen, X.; Huang, K. MemVLT: Vision-language tracking with adaptive memory-based prompts. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 14903–14933.
13. Hou, X.; Xing, J.; Qian, Y.; Guo, Y.; Xin, S.; Chen, J.; Tang, K.; Wang, M.; Jiang, Z.; Liu, L.; et al. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 26551–26561.
14. Alansari, M.; Javed, S.; Ganapathi, I.I.; Alansari, S.; Naseer, M. CLDTracker: A Comprehensive Language Description for Visual Tracking. *Inf. Fusion* **2025**, *124*, 103374. [CrossRef]
15. Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-efficient transfer learning for NLP. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 2790–2799.
16. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *ICLR* **2022**, *1*, 3.

17. Lester, B.; Al-Rfou, R.; Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv* **2021**, arXiv:2104.08691. [CrossRef]
18. Yang, J.; Li, Z.; Zheng, F.; Leonardis, A.; Song, J. Prompting for multi-modal tracking. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 3492–3500.
19. Zhu, J.; Lai, S.; Chen, X.; Wang, D.; Lu, H. Visual prompt multi-modal tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 9516–9526.
20. Borsuk, V.; Vei, R.; Kupyn, O.; Martyniuk, T.; Krashenyi, I.; Matas, J. FEAR: Fast, efficient, accurate and robust visual tracker. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 644–663.
21. Huang, L.; Zhao, X.; Huang, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1562–1577. [CrossRef] [PubMed]
22. Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; Felsberg, M. Eco: Efficient convolution operators for tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6638–6646.
23. Bhat, G.; Danelljan, M.; Gool, L.V.; Timofte, R. Learning discriminative model prediction for tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 6182–6191.
24. Xie, F.; Wang, Z.; Ma, C. Diffusiontrack: Point set diffusion model for visual object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 19113–19124.
25. Siméoni, O.; Vo, H.V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M.; et al. Dinov3. *arXiv* **2025**, arXiv:2508.10104. [PubMed]
26. Chen, X.; Wu, Z.; Liu, X.; Pan, Z.; Liu, W.; Xie, Z.; Yu, X.; Ruan, C. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv* **2025**, arXiv:2501.17811. [CrossRef]
27. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. Dinov2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
28. Fang, Y.; Sun, Q.; Wang, X.; Huang, T.; Wang, X.; Cao, Y. Eva-02: A visual representation for neon genesis. *Image Vis. Comput.* **2024**, *149*, 105171. [CrossRef]
29. Botach, A.; Zheltonozhskii, E.; Baskin, C. End-to-end referring video object segmentation with multimodal transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4985–4995.
30. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PmLR, Online, 18–24 July 2021; pp. 8748–8763.
31. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, PmLR, Sydney, Australia, 6–11 July 2015; pp. 448–456.
32. Pydimarry, S.A.; Khairnar, S.M.; Palacios, S.G.; Sankaranarayanan, G.; Hoagland, D.; Nepomnayshy, D.; Nguyen, H.P. Evaluating model performance with hard-swish activation function adjustments. *arXiv* **2024**, arXiv:2410.06879. [CrossRef]
33. Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; Douze, M. Levit: A vision transformer in convnet's clothing for faster inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Online, 10–17 October 2021; pp. 12259–12269.
34. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Republic of Korea, 27–28 October 2019; pp. 658–666.
35. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
36. Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; Ghanem, B. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 300–317.
37. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
38. Zheng, Y.; Zhong, B.; Liang, Q.; Li, G.; Ji, R.; Li, X. Toward unified token learning for vision-language tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 2125–2135. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.