KEYWORD ENHANCED GENERATIVE PRIORS FOR VISION LANGUAGE TRACKING IN UNDERWATER CONDITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

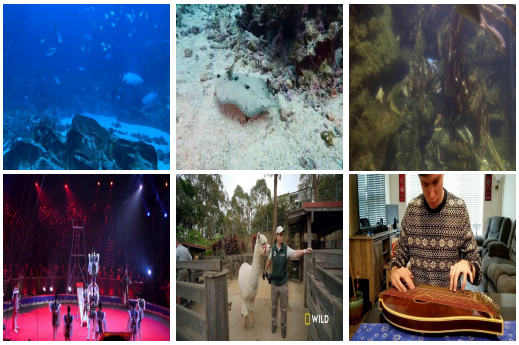
Vision Language Tracking (VLT) localizes target objects in videos by jointly exploiting visual and textual cues, ensuring stronger reliability than traditional tracking methods. However, existing VLT methods still suffer from severe performance degradation in underwater scenarios because the strong noise and color distortion resemble the target and its surroundings closely. Additionally, the lack of corresponding training data prevents this limitation from being addressed through a pure data-driven approach. In this paper, we address this limitation through a more efficient method, transferring the strong prior of generative models learned from the pretraining on large-scale natural images to underwater scenarios. Specifically, we incorporate a keyword enhancement strategy. It extracts intermediate multi-modal features from the generation process to assist tracking. During this, it extracted the keywords related to the target and modifies cross-attention maps by increasing the corresponding attention weights to steer the features to place greater emphasis on the tracking target. With this in mind, we propose KeyTrack, a generative-model-based VLT framework, which uses strong priors from generative models and enables target-focused representation learning while suppressing irrelevant environmental noise, thereby achieving more accurate tracking. Extensive experiments demonstrate that KeyTrack achieves state-of-the-art performance on underwater tracking benchmarks and maintains competitive results on general datasets, validating its robustness and generalization capability.

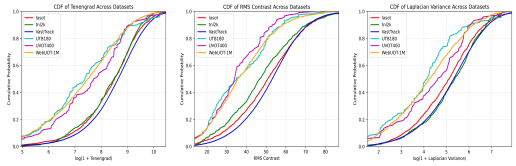
1 INTRODUCTION

Vision Language Tracking (VLT) aims to locate a target in video by jointly leveraging visual and textual cues. By integrating language semantic with visual appearance (Zheng et al., 2023), VLT reduces target ambiguity and improves robustness compared to conventional visual trackers (Bertinetto et al., 2016; Bhat et al., 2018; Danelljan et al., 2019; Song et al., 2018; Wang et al., 2021b; Chen et al., 2023b; Wei et al., 2023) that rely solely on appearance features. This multi-modal tracker enables consistent localization under challenging conditions and has been applied in real-world scenarios (Feng et al., 2021; Wang et al., 2021a; Zhou et al., 2023; Guo et al., 2022), including video surveillance, robotics, and autonomous driving. As a result, VLT has attracted increasing attention in underwater environments, where it holds significant potential for applications such as marine life monitoring (Cai et al., 2023), underwater search and rescue (Alawode et al., 2023), underwater imaging (Li et al., 2023b), and maritime security (González-Sabbagh & Robles-Kelly, 2023).

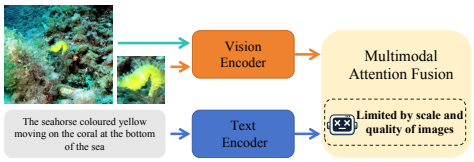
However, current VLT models in Figure 1b face numerous challenges in underwater environments. As illustrated in Figure 1a, compared with common images, issues—such as uneven lighting (Zhang et al., 2025a), low visibility, low contrast, water-induced color shifts, visually similar distractors, camouflage, and motion blur—are more prevalent in underwater images. Quantitative analysis based on Cumulative Distribution Function (CDF) curves across several tracking benchmarks shows that underwater datasets are concentrated in lower-value regions for Tenengrad gradient, Laplacian variance, and Root Mean Square (RMS) contrast, indicating inferior visual quality. Together, they corroborate image degradation and noise, which makes underwater tracking substantially more challenging than common tracking (Zhang et al., 2022; Fan et al., 2019; Huang et al., 2019; Wang et al., 2021a; Muller et al., 2018; Zhang et al., 2024a). In addition, behavioral patterns of marine life and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

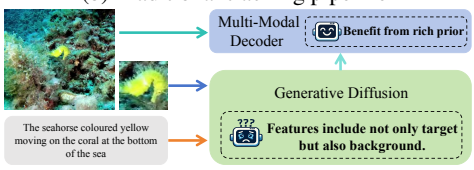




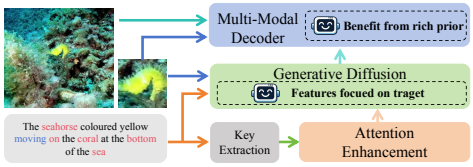
(a) Dataset analysis.



(b) Traditional tracking pipeline



(c) Diffusion tracking pipeline



(d) Our tracking pipeline

Figure 1: (a) Image samples and quality analysis of underwater and conventional tracking datasets. The first row shows underwater tracking examples, the second row shows conventional tracking examples, and the third row presents image quality analysis using Tenengrad, RMS contrast, and Laplacian variance. Curves further to the right indicate higher image quality. (b) Traditional tracking pipeline. (c) Diffusion-based tracking pipeline. (d) Our tracking pipeline built upon the proposed keyword enhancement strategy.

scarcity of large-scale underwater datasets further limit effectiveness of existing trackers (Alawode et al., 2023; Cai et al., 2023). Consequently, there is an urgent need to develop robust underwater tracking methods that directly improve performance in underwater conditions.

To address the challenges encountered in underwater scenarios, we introduce generative models into tracking, as shown in Figure 1c. They are typically trained on large-scale datasets, thus providing rich visual priors. Guided by discriminative target-specific textual descriptions, these priors can alleviate underwater data scarcity to some extent. Moreover, previous work (Luo et al., 2026) shows that generative models can fuse textual and visual cues, providing rich semantic information (Baranchuk et al., 2021) and high-quality vision–language representations, thus enabling models to better capture target appearance and motion. Consequently, they can refine or restore visual templates during tracking, which has the potential to mitigate image degradation and noise (Whang et al., 2022; Rahman et al., 2023), thereby enhancing feature representations and overall tracking performance.

However, since the generative models are originally designed to reconstruct the entire image rather than a specific target. So they still reconstruct background regions unrelated to the target as well, even when guided by discriminative target-specific textual descriptions. In applications, the features obtained during the restoration stage capture not only the tracked object but also environmental background and other noise. This issue becomes more pronounced in underwater scenarios, where targets often exhibit high visual similarity to the background. Such global enhancement may suppress discriminative target cues or over-amplify background regions, leading to feature representations that are biased toward the surrounding environment rather than the tracked object.

To address this limitation, we analyze generative models and find that the provided textual guidance is not solely focused on the target, but also on the environment. As shown in Figure 2, t-SNE shows that the features related to the target and environment exhibit significant differences. Meanwhile, although the generation process benefits from textual guidance and shows stronger responses to the target, considerable attention is still allocated to background regions, which is particularly evident for non-keyword tokens. While effective for generation, this behavior introduces background-biased features that are unrelated to the target and different from target features, thus degrading accuracy.

Accordingly, we propose a keyword enhancement strategy that extracts salient words from descriptions and boosts target importance during generative restoration and feature extraction. By reinforcing target semantics, it guides the generative model to prioritize target repair and encoding while



Figure 2: Visualizations of the template image, the t-SNE of vision features, the cross-attention map for keyword and non-keyword tokens.

suppressing background interference. This target-aware enhancement provides discriminative and reliable representations for underwater tracking.

Based on the above analysis, we propose KeyTrack demonstrated in Figure 1d, a Vision Language Tracking framework built upon a generative model and centered on the proposed keyword enhancement strategy. By leveraging the rich prior knowledge of generative models, KeyTrack effectively mitigates image degradation and noise in underwater environments. Meanwhile, we enforce target-focused feature extraction through enhanced semantic cues, and enables more accurate target localization, leading to improved tracking performance.

Experimental results show that our proposed KeyTrack performs favorably on underwater datasets such as WebUOT-1M (Zhang et al., 2024a), UW-COT220 (Zhang et al., 2024b) and UVOT400 (Alawode et al., 2023), outperforming existing tracking approaches. Moreover, it maintains competitive performance on general benchmarks including LaSOT (Fan et al., 2019), TNL2K (Wang et al., 2021a) and MGIT (Hu et al., 2024), demonstrating strong robustness and generalization capability. Our main contributions are summarized as follows:

1. We propose a keyword-guided enhancement strategy to better exploit rich generative visual priors, reducing background bias and improving target-aware feature extraction in challenging underwater scenarios.
2. We propose KeyTrack, a Vision Language Tracking framework with generative backbone and keyword enhancement for target-focused multi-modal feature extraction.
3. Experimental results on underwater and general tracking benchmarks show outstanding performance, validating its effectiveness, robustness, and generalization.

2 Related Work

Vision Language Tracking. Vision Language Tracking (VLT) localizes targets in video by exploiting visual and language information (Li et al., 2017; Feng et al., 2020; Yang et al., 2020; Li et al., 2022; Guo et al., 2022; Shao et al., 2024). Unlike visual trackers relying only on appearance cues, VLT incorporates language as semantic supervision, reducing ambiguities from similar distractors and improving robustness under occlusion, illumination variation, and complex backgrounds. A core challenge is effective fusion of visual and textual modalities. To address this, existing methods (Feng et al., 2021; Wang et al., 2021a) adopt pre-trained language models, such as BERT (Devlin et al., 2018) or GPT (Brown et al., 2020), to encode descriptions into dense embeddings that are integrated with visual features through different mechanisms. Owing to their ability to model cross-modal dependencies, Transformer-based architectures (Vaswani et al., 2017) have become dominant for multi-modal fusion in VLT, with methods such as TransVLT (Zhao et al., 2023) and JointNLT (Zhou et al., 2023) aligning visual and linguistic features through cross-attention.

Underwater Object Tracking. Underwater object tracking (UOT) (Luo et al., 2021; Panetta et al., 2021; Cai et al., 2023; Alawode et al., 2023; 2022) aims to estimate object locations in diverse underwater scenes and has gained attention in marine ecology and conservation. Early methods relied on handcrafted pipelines; for example, Yoerger et al. (Yoerger et al., 2021) introduced the Mesobot platform to track marine organisms using color segmentation and blob tracking (Bradski, 2000), effective mainly in constrained environments. Later studies explored learning-based paradigms, in-

cluding tracking-by-detection with deep neural networks (Katija et al., 2021) and hybrid training strategies addressing data imbalance and model drift (Li et al., 2023b). Several UOT benchmarks have since been introduced (Luo et al., 2021; Panetta et al., 2021; Alawode et al., 2022; Cai et al., 2023; Alawode et al., 2023), but most remain limited in scale and diversity, restricting robust and generalizable model learning. Although WebUOT-1M (Zhang et al., 2024a) partly alleviates this issue, data scarcity and domain bias still hinder underwater tracking.

Generative Model. Diffusion models (Ho et al., 2020; Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2020; Lugmayr et al., 2022) have emerged as powerful generative approaches in computer vision, using progressive denoising to reconstruct clean samples from corrupted inputs, as demonstrated by DDPMs (Ho et al., 2020; Sohl-Dickstein et al., 2015), score-based models (Song & Ermon, 2019; 2020), and Score SDEs (Karras et al., 2022; Song et al., 2021; 2020). With large-scale Vision Language models, diffusion techniques have extended to cross-modal generation, exemplified by Stable Diffusion (Rombach et al., 2022) and DALLE 2 (Rassin et al., 2022), enabling high-fidelity text-conditioned synthesis. Building on this capability, diffusion-based frameworks have been applied to image restoration and representation enhancement, improving texture reconstruction and semantic consistency (Li et al., 2025b; Zhao et al., 2020) and providing reliable features for downstream vision tasks (Russakovsky et al., 2015; Quan et al., 2024). Recent studies show that integrating diffusion models into enhancement or denoising pipelines provides structural and semantic guidance, benefiting object detection (Chen et al., 2023a; Pnvr et al., 2023; Cheng et al., 2025; Ge et al., 2025) and improving visual representation robustness.

3 METHODOLOGY

Benefiting from the observation that considerable attention is allocated to background regions as well, which is particularly evident for non-keyword tokens, we design KeyTrack shown in Figure 3, a generative model centered on keyword enhancement strategy. Given a template along with its text, Key Extraction Module first identifies the core keywords related to the target from the textual input and feeds them into Attention Enhancement Module, which guides Generative Diffusion Module to perform target-aware multi-modal feature extraction. Vision encoder then encodes search and template image and concatenate them to obtain visual features. The extracted multi-modal features and visual features are subsequently integrated in Multi-Modal Decoder, generating fused representations that are finally passed to prediction head to produce the final predicted bounding box. We will introduce each module in the following sections.

3.1 KEY EXTRACTION MODULE

Since different words induce different attention tendencies, we first need to select keywords from the text that are more closely related to the tracking target. To this end, we design the Key Extraction Module. The primary objective of this Module is to identify words in the textual description that are strongly correlated with the tracking target, such as category-related and action-related terms. Compared to contextual expressions, target-related words are generally more stable and can provide reliable semantic cues throughout the tracking process. However, due to the flexibility and diversity of natural language, accurately distinguishing target-specific keywords from auxiliary contextual words remains challenging. Existing Vision Language Tracking methods often rely on the similarity of targets and text to implicitly identify such keywords, which introduces a strong dependence on fine-grained multi-modal alignment and increases optimization and deployment difficulty.

To address this issue, we decouple keyword identification from visual information and perform extraction purely in the textual domain. Following prior work (Feng et al., 2025a), keyword extraction is formulated as a multi-label binary classification task, where each token is categorized as target-related (to be enhanced) or context-related (non-enhanced). We adopt the keyword annotations in (Feng et al., 2025a) as supervision and learn keywords extraction using a lightweight Multi Layer Perceptron (MLP) built on the text features f_{text}:

$$p_T = \mathrm{MLP}(f_{text}), \tag{1}$$

where $p_T \in [0,1]^{N_l \times 1}$ denotes the probability of each token being a target keyword. The original text features are then reweighted by $p_T > 0.5$ to obtain the token locations $\mathcal{L}$ that require enhancement. By explicitly modeling target-related tokens in this manner, the proposed Key Extraction

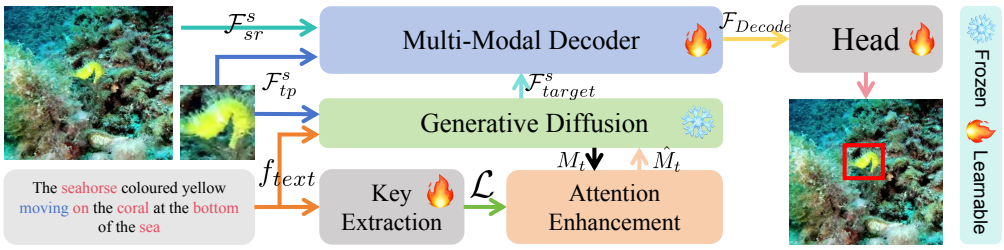


Figure 3: Overview of KeyTrack. The text is first processed by Key Extraction Module for keyword extraction. The text and template are then generatively fused under the guidance of Attention Enhancement Module using Generative Diffusion Module. Finally, search and template features are fused with multi-modal features from generative model for tracking.

Module enables the identification of words in the textual description that are strongly correlated with the tracking target, providing a concise and reliable textual prior. This target-aware representation serves as the foundation for subsequent attention enhancement and target-aware multi-modal feature extraction in Vision Language Tracking.

3.2 Attention Enhancement Module

Once target-related keywords are identified, they can be explicitly used to guide attention redistribution. This defines the objective of the Attention Enhancement Module, as shown in Figure 4. Its goal is to amplify attention responses associated with the tracking target while suppressing irrelevant background regions. In this way, it guides the Generative Diffusion Module to extract target-aware features that are better suited for Vision Language Tracking. Formally, we denote

$$z_{t-1}, M_t, \mathcal{F}_t^s = \mathrm{DM}(z_t, f_{text}, t, s) \tag{2}$$

as a diffusion step at timestep t, where z_t represents the noisy latent representation, f_{text} denotes the textual prompt features and s indicates the random seed. This operation outputs the denoised latent z_{t-1} together with the attention map M_t and intermediate features $\mathcal{F}_t^s$. We further define

$$z_{t-1}, \hat{M}_t, \mathcal{F}_t^s = \mathrm{DM}(z_t, f_{text}, t, s)\{M_t \leftarrow \hat{M}_t\} \tag{3}$$

as a modified diffusion step, in which the original attention map M_t is replaced with an enhanced attention map $\hat{M}_t$, while the value embeddings V conditioned on the prompt remain unchanged, as shown in Figure 4. This formulation allows attention responses to be selectively manipulated without altering the underlying semantic content encoded by the generative model. Specifically, the enhanced attention map $\hat{M}_t$, which emphasizes the target-related regions, is obtained through a general attention manipulation function as below:

$$\hat{M}_t = \mathrm{Edit}(M_t, \mathcal{L}), \tag{4}$$

which takes the attention map M_t and token locations $\mathcal{L}$ corresponding to keywords. This function increases the attention weights associated with keywords while attenuating responses from non-target regions, thereby steering the diffusion process toward more target-focused feature extraction. During tracking, within the Generative Diffusion Module guided by the Attention Enhancement Module, the diffusion process is performed iteratively using the original textual description, while the attention map is adjusted in all cross-attention layers at each step according to the keyword locations. To ensure stable and consistent feature enhancement, the internal randomness of the diffusion process is fixed across steps.

Algorithm 1 outlines the pipeline of the proposed attention enhancement strategy. The key operation lies in manipulating attention maps during the diffusion process, which is realized through the attention editing function $\mathrm{Edit}(\cdot)$. Speicially, Let $\widehat{M}_t \in \mathbb{R}^{N_q \times N_k}$ denote the attention map at diffusion step t, where N_q and N_k correspond to the numbers of query and key tokens, respectively, and let $\mathcal{L}$ represent the index set of tokens associated with the extracted target keywords. To strengthen target-related semantics while suppressing background interference, we selectively rescale the attention responses corresponding to these keyword tokens. Formally, thisoperation is defined as

$$(\mathrm{Edit}(M_t, \mathcal{L}))_{i,j} = \begin{cases} c \cdot (M_t)_{i,j}, & \text{if } j \in \mathcal{L}, \\ (M_t)_{i,j}, & \text{otherwise}, \end{cases} \tag{5}$$

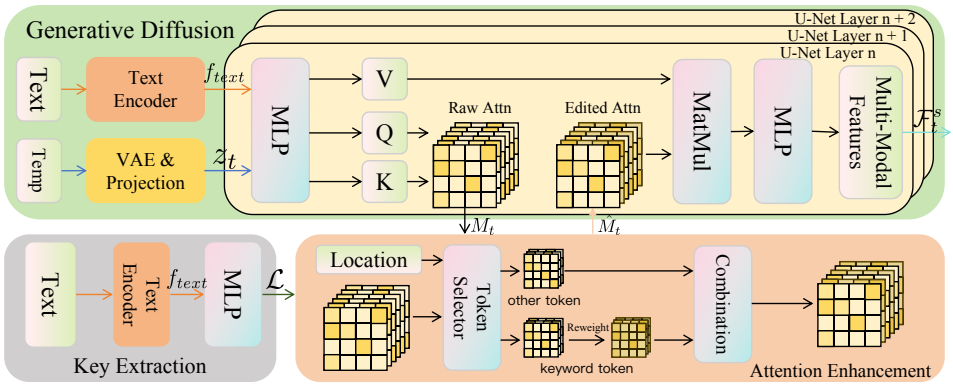


Figure 4: Pipeline of the proposed Key Enhancement Strategy. Keywords are first extracted from text in Key Extraction Module. During cross-attention, Attention Enhancement Module reweights the attention of the corresponding tokens based on the identified keywords, thus guiding the process and make the features more focused on the tracking target.

where i and j index the query and key dimensions of the attention map, respectively, and $c > 1$ is a scaling factor controlling the strength of enhancement. By explicitly amplifying target-related tokens, this formulation promotes tighter alignment between textual semantics and the tracking target throughout the diffusion process. In practice, the scaling factor $c = 1.1$ is fixed across diffusion steps to maintain stable and consistent feature modulation. Incorporating this attention editing mechanism into each denoising step enables the generative model to leverage its learned priors in a target-aware manner, resulting in target-aware multi-modal features $\mathcal{F}^s_{target} = \mathcal{F}^s_1$ for downstream tracking.

Algorithm 1: Attention Enhancement Strategy

Input: Text features f_{text}, keyword locations $\mathcal{L}$, random seed s.
Output: Enhanced latent representation z_0 and Target-aware Multi-modal features $\mathcal{F}^s_1$.
Initialize $z_T \sim \mathcal{N}(0, I)$ with random seed s;
for $t = T, T-1, \ldots, 1$ **do**
  $M_t \leftarrow \mathrm{DM}(z_t, f_{text}, t, s)$;
  $\widehat{M}_t \leftarrow \mathrm{Edit}(M_t, \mathcal{L})$;
  $(z_{t-1}, \mathcal{F}^s_t) \leftarrow \mathrm{DM}(z_t, f_{text}, t, s)\{M \leftarrow \widehat{M}_t\}$;
end
Return $z_0, \mathcal{F}^s_1$;

Overall, this mechanism explicitly steers the generative model toward target-aware optimization instead of global image restoration, thereby effectively reducing background bias during feature restoration and yielding more discriminative representations for Vision Language Tracking.

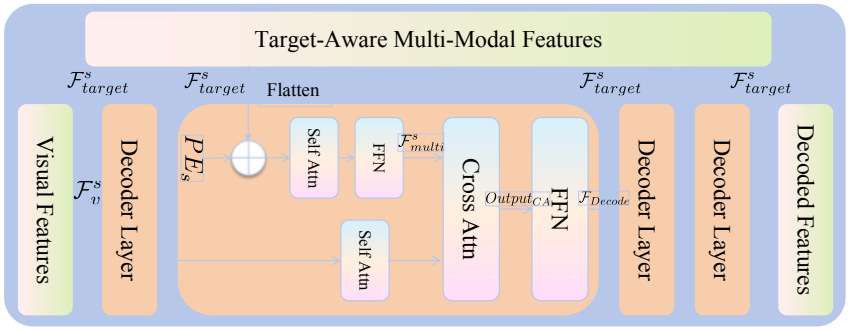


Figure 5: Multi-Modal Decoder Architecture. It fuses target-aware multimodal features and visual features to generate fused representations.

3.3 Multi-Modal Decoder

After obtaining the target-aware multi-modal features $\mathcal{F}^s_{target}$, we fuse them with the visual features $\mathcal{F}^s_v = \text{Concat}(\mathcal{F}^s_{tp}, \mathcal{F}^s_{sr})$ encoded by the vision encoder through a Multi-Modal Decoder, as shown in Fig. 5. Specifically, the target-aware multi-modal features are first flattened and enhanced with positional embeddings, while the visual features are processed in parallel. Both streams are refined by separate self-attention layers to strengthen their contextual representations. The multi-modal stream is then aligned by an FFN and used to interact with the visual stream through cross-attention, enabling target-aware fusion between language-enhanced cues and visual search features. Finally, another FFN consolidates the attended features, and the decoded representation is forwarded to the prediction head in the last decoder layer for target localization, as illustrated in Fig. 3.

4 Experiments

4.1 Experiment Settings

KeyTrack is implemented using Python 3.8.18 and PyTorch 2.1.1. Training and test processes are conducted on 4 Nvidia Quadro RTX A800. We adopt Stable Diffusion v1.5 for generative multimodal fusion and the pretrained Fast-iTPN-B (Tian et al., 2024) to extract high-quality visual features. We train the model under underwater and conventional settings. It is first trained for 200 epochs on LaSOT (Fan et al., 2019), OTB99-Lang (Li et al., 2017), TNL2K (Wang et al., 2021a), TrackingNet (Muller et al., 2018), VastTrack (Peng et al., 2024), and RefCOCOg (Mao et al., 2016). During the next 100 epochs, the underwater setting additionally uses WebUOT-1M (Zhang et al., 2024a), while the conventional setting retains the original datasets. We use AdamW (Loshchilov & Hutter, 2017) with a weight decay of 10^{-4}, gradient clipping of 0.1, and a learning rate warmed up to 4×10^{-4} over 30 epochs and decayed to 4×10^{-5} after epoch 240. The compound loss includes L1 and GIoU losses for box regression and a BCE loss for keyword supervision:

$$L_{loc} = \lambda_{L1} L_1(B_i, \hat{B}_i) + \lambda_{giou} L_{giou}(B_i, \hat{B}_i) + \lambda_{bce} L_{bce}, \tag{6}$$

where $\lambda_{L1} = 5$, $\lambda_{bce} = 0.2$ $\lambda_{giou} = 2$ are loss weights, B_i is the ground-truth box and $\hat{B}_i$ is predicted target box. During evaluation, we follow standard tracking practice by applying a Hanning window penalty to the prediction outputs, incorporating positional priors.

Table 1: Comparison on Underwater datasets. **red** : best; **blue**: second best.

Method	WebUOT-1M			UW-COT220		UTB180		UVOT400		Speed
	$AUC(\%)$	$P_{Norm}(\%)$	$P(\%)$	$AUC(\%)$	$P(\%)$	$AUC(\%)$	$P(\%)$	$PR(\%)$	$SR(\%)$	FPS
HIPTrack (Cai et al., 2024)	61.6	-	58.0	57.9	53.6	66.1	60.7	-	-	**92.1**
MixFormer (Cui et al., 2022)	-	-	45.6	-	-	57.4	50.6	62.5	63.5	-
CiteTracker (Li et al., 2023a)	54.6	58.4	49.3	-	-	61.7	54.5	61.9	-	-
UOSTrack (Li et al., 2023b)	61.1	-	54.2	56.9	50.9	66.7	60.3	59.6	**66.5**	70.1
OKTrack (Zhang et al., 2024a)	60	-	57.5	59.2	53.8	**70.1**	**67.3**	58.4	63.2	69.8
SeqTrack (Chen et al., 2023b)	54.0	58.5	50.8	56.0	50.6	60.8	55.2	63.2	65.4	28.1
GRM (Gao et al., 2023)	58.9	-	55.3	56.1	50.1	64.7	59.2	**64.5**	65.5	35.2
ARTrack (Wei et al., 2023)	**63.9**	-	57.9	60.0	55.9	69.0	65.1	61.3	65.0	37.3
ODTrack (Zheng et al., 2024)	63.8	**73.6**	**59.1**	**60.4**	**56.4**	**70.1**	64.6	-	-	34.4
CTIFTrack (Zhang et al., 2025b)	60.8	-	63.4	-	-	-	-	64.3	**65.9**	-
USBTrack (Xu et al., 2025)	63.2	-	63.8	-	-	67.9	63.1	-	-	-
SUTrack (Chen et al., 2025)	61.5	-	62.6	-	-	-	-	64.5	65.3	-
ORTrack (Wu et al., 2025)	52.2	-	45.1	45.2	35.2	56.7	47.2	-	-	**109.7**
KeyTrack (Ours)	**64.2**	**75.6**	**64.0**	**63.3**	**61.8**	**70.6**	**71.1**	**65.0**	**66.5**	65.2

4.2 Comparison on Underwater Tracking Datasets

We evaluate on WebUOT-1M (Zhang et al., 2024a), UW-COT220 (Zhang et al., 2024b), UTB180 (Alawode et al., 2022), and UVOT400 (Alawode et al., 2023). WebUOT-1M is the first million-scale underwater tracking benchmark, covering diverse categories and challenging conditions such as low resolution and visibility degradation, with prompts for multi-modal evaluation. UW-COT220 focuses on camouflaged object tracking, including 220 videos from 96 categories where targets are highly similar to underwater backgrounds. UTB180 is a traditional underwater tracking benchmark with accurate annotations and diverse challenges such as motion blur and occlusion. UVOT400 is an underwater long-term tracking benchmark involving similar objects and frequent occlusions, requiring robust global and local feature modeling. As shown in Tab. 1, our

method consistently improves over prior trackers: on WebUOT-1M, it outperforms OKTrack (Zhang et al., 2024a) by 4.2% AUC and 6.5% Precision; on UW-COT220, it surpasses ODTrack (Zheng et al., 2024) by 2.9% AUC and 5.4% Precision; on UTB180, it improves over ODTrack by 0.5% AUC and 6.5% Precision; and on UVOT400, it exceeds GRM (Gao et al., 2023) by 0.5% AUC and 1.0% Precision. Speed is critical for practical tracking, as high inference efficiency enables prompt responses to dynamic underwater changes. As shown in Tab. 1, KeyTrack runs at 65.2 FPS, comparable to OKTrack (Zhang et al., 2024a), while maintaining high accuracy. These results demonstrate efficient and accurate underwater tracking, further validating the effectiveness of our method.

Table 2: Comparison on Traditional datasets. **red** : best; **blue**: second best.

Method	LaSOT			$\text{LaSOT}_{\text{ext}}$		TNL2K			MGIT		
	$AUC(\%)$	$P_{Norm}(\%)$	$P(\%)$	$AUC(\%)$	$P(\%)$	$AUC(\%)$	$P_{Norm}(\%)$	$P(\%)$	$AUC(\%)$	$P_{Norm}(\%)$	$P(\%)$
QueryNLT (Shao et al., 2024)	59.9	69.6	63.5	-	-	57.8	75.6	58.7	-	-	-
MMTrack (Zheng et al., 2023)	70.0	82.3	75.7	49.4	55.3	58.6	75.2	59.4	-	-	-
OneTracker (Hong et al., 2024)	70.5	79.9	76.5	-	-	58.0	-	59.1	-	-	-
UVLTrack (Ma et al., 2024)	69.4	-	74.9	49.2	55.8	62.7	-	65.4	-	-	-
CTVLT (Feng et al., 2025b)	72.3	-	79.7	-	-	62.2	-	79.5	69.2	-	62.9
MemVLT (Feng et al., 2024)	72.9	**85.7**	**80.5**	52.1	59.8	63.3	80.9	67.4	**69.4**	**81.3**	**63.7**
DMTrack (Zhang et al., 2024c)	66.8	-	72.7	47.3	52.1	57.7	-	59.9	-	-	-
GLAD (Luo et al., 2026)	69.5	79.6	74.2	50.6	55.7	59.7	77.6	62.2	-	-	
SUTrack (Chen et al., 2025)	**73.2**	83.4	**80.5**	**53.1**	**60.5**	**65.0**	-	67.9	-	-	-
DUTrack (Li et al., 2025a)	**73.0**	**83.8**	**81.1**	50.5	58.1	64.9	**82.9**	**70.6**	-	-	-
KeyTrack (Ours)	**73.0**	82.6	79.6	**53.3**	**61.5**	**66.4**	**84.2**	**71.9**	**72.9**	**84.0**	**68.2**

4.3 Comparison on Traditional Tracking Datasets

To evaluate generalization beyond underwater scenarios, we further test our method on conventional vision-language tracking benchmarks, including LaSOT (Fan et al., 2019), TNL2K (Wang et al., 2021a), $\text{LaSOT}_{\text{ext}}$ (Fan et al., 2021), and MGIT (Hu et al., 2024). LaSOT and $\text{LaSOT}_{\text{ext}}$ extend visual tracking benchmarks with language annotations and focus on long-term tracking scenarios (Fan et al., 2019; Li et al., 2017), while $\text{LaSOT}_{\text{ext}}$ further introduces many distractors. TNL2K is a large-scale VLT benchmark with over two thousand sequences, diverse targets, and fine-grained text annotations for evaluating vision-language interaction (Wang et al., 2021a). MGIT is a recent VLT benchmark designed for complex long-term scenarios with multi-granularity language annotations. As shown in Tab. 2, our method achieves performance comparable to SUTrack (Chen et al., 2025) on LaSOT and outperforms it on $\text{LaSOT}_{\text{ext}}$ by 0.2% and 1.0% across key metrics. On TNL2K, it improves over DUTrack (Li et al., 2025a) by 1.5% AUC and 0.7% Precision. On MGIT, it surpasses MemVLT (Feng et al., 2024) by 3.5%, 2.7%, and 4.5% in AUC, PNorm, and Precision, respectively. These results show that our method maintains strong performance on conventional benchmarks, demonstrating good generalization capability beyond underwater tracking scenarios.

4.4 Analysis about the Keyword Scaling Factor

To evaluate the proposed Key Enhancement Strategy, we conduct an ablation study by varying the keyword scaling factor c on TNL2K (Wang et al., 2021a) and WebUOT-1M (Zhang et al., 2024a), where $c = 1$ denotes the baseline without the module. As reported in Table 3, as c increases, the performance initially improves as the scaling factor increases, reaches its peak at $c = 1.1$ and then gradually declines when the scaling factor becomes larger: moderate scaling strengthens target-related features while suppressing background noise, leading to improved multi-modal fusion and tracking accuracy, which verifies the effectiveness of our proposed method. However, excessive scaling over-suppresses environmental cues, diminishing contextual information and ultimately harming generalization and overall tracking performance. Therefore, we adopt $c = 1.1$ finally.

Table 3: Exploration studies. c denotes the enhancement coefficient, with $c = 1$ indicating no enhancement. Parentheses show differences from our adopted setting ($c = 1.1$).

Methods	WebUOT-1M			TNL2K		
	$AUC(\%)$	$P_{Norm}(\%)$	$P(\%)$	$AUC(\%)$	$P_{Norm}(\%)$	$P(\%)$
KeyTrack($c = 1$)	63.5 (-0.7)	74.6 (-1.0)	63.4 (-0.6)	65.4 (-1.0)	82.7 (-1.5)	70.1 (-1.8)
KeyTrack($c = 1.05$)	63.8 (-0.4)	75.0 (-0.6)	63.3 (-0.7)	65.6 (-0.8)	83.4 (-0.8)	71.2 (-0.7)
KeyTrack($c = 1.1$)	**64.2**	**75.6**	**64.0**	**66.4**	**84.2**	**71.9**
KeyTrack($c = 1.15$)	63.4 (-0.8)	74.8 (-0.8)	63.0 (-1.0)	65.8 (-0.6)	84.0 (-0.2)	71.6 (-0.3)
KeyTrack($c = 1.2$)	62.1 (-2.1)	74.2 (-1.4)	62.3 (-1.7)	65.2 (-1.2)	82.8 (-1.4)	70.3 (-1.6)

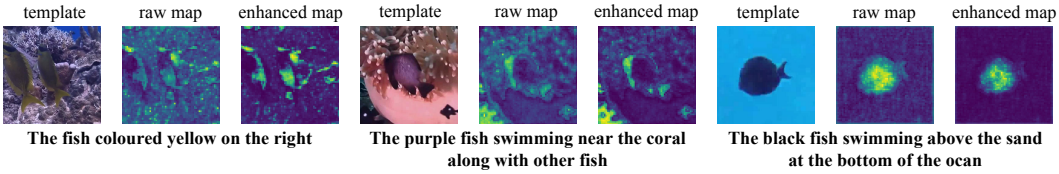


Figure 6: Visualizations of Key Enhancement Strategy.

4.5 Attention Map Visualization

To qualitatively analyze the impact of the proposed Key Enhancement Strategy, we visualize the average cross-attention maps generated during the diffusion process. As illustrated in Figure 6, the attention corresponding to target-related regions such as fish and Caribbean tang are more strongly activated and receive higher attention scores, while irrelevant environmental background regions such as coral and sand receive lower attention scores. These visualizations demonstrate that our Key Enhanced Strategy effectively guides feature generation and makes the multi-modal features more focusing on the tracking target, thereby better supporting tracking process.

4.6 Robustness Analysis

Herein, we discuss the limitations of our method. Since it relies on keyword extraction, accurate textual keywords are required. Under normal conditions, the Key Extraction Module achieves 98% accuracy on the test set, effectively avoiding severe failure cases. However, textual descriptions may be ambiguous or incorrect in practice. To evaluate robustness under erroneous text, we conduct additional experiments, with results shown in Figure 7. When the scene is relatively simple, the controlled keyword scaling factor c=1.1 prevents incorrect textual cues from excessively dominating feature extraction, allowing successful target localization despite erroneous descriptions. In more challenging scenarios with visually similar objects, incorrect text may occasionally mislead tracking and cause failure. These results indicate that template visual features, which directly interact with the search image in the Multi-Modal Decoder, provide an additional grounding signal. This complementary design establishes a lower performance bound: although erroneous text may reduce tracking accuracy, the target can still be successfully localized in most cases.

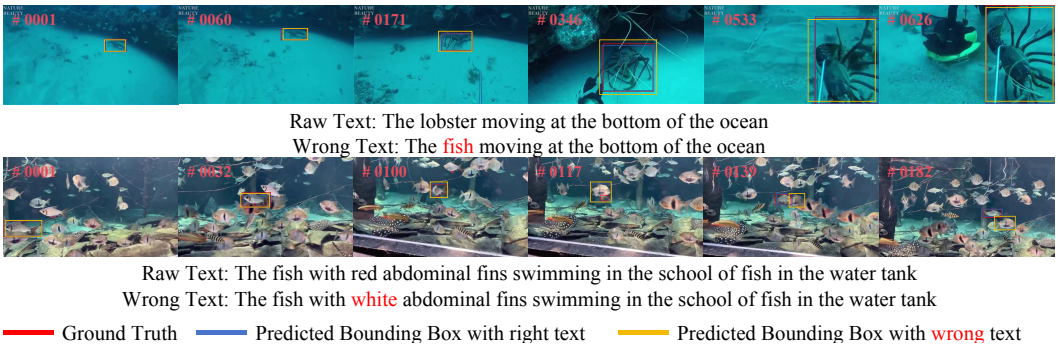


Figure 7: Analysis about the effect of wrong natural language. The figure above shows an example of successful tracking, while the one below illustrates a failure case.

5 Conclusion

We present KeyTrack, a Vision Language Tracking framework designed for challenging underwater environments. To address the strong environment noise and the scarcity of large-scale underwater datasets, KeyTrack transfers the strong prior of generative models and integrates a Key Enhancement Strategy that emphasizes target-relevant features through keyword extraction and attention-guided feature fusion. Experiments on underwater benchmarks show clear improvements over existing methods, while results on conventional datasets demonstrate the strong generalization. Ablation studies and cross-attention visualizations further confirm the effectiveness of the proposed design enabling robust and accurate tracking in complex underwater scenarios.

AI USE STATEMENT

The authors used generative AI tools to assist with language polishing, grammar checking,. All AI-assisted content was reviewed and revised by the authors. The authors take full responsibility for the final content of the paper, including all text, claims, experiments, and code.

ETHICS STATEMENT

This work uses publicly available datasets for visual tracking research and does not involve human subject experiments, private personal data, or sensitive attributes. We do not foresee direct negative societal impacts beyond the general dual-use considerations associated with visual tracking technologies.

REPRODUCIBILITY STATEMENT

We provide implementation details, training settings, dataset configurations, evaluation protocols, and ablation studies in the main paper and appendix. Code and configuration files will be released upon publication to facilitate reproducibility.

REFERENCES

Basit Alawode, Yuhang Guo, Mehnaz Ummar, Naoufel Werghi, Jorge Dias, Ajmal Mian, and Sajid Javed. Utb180: A high-quality benchmark for underwater tracking. In *Proceedings of the Asian conference on computer vision*, pp. 3326–3342, 2022.

Basit Alawode, Fayaz Ali Dharejo, Mehnaz Ummar, Yuhang Guo, Arif Mahmood, Naoufel Werghi, Fahad Shahbaz Khan, Jiri Matas, and Sajid Javed. Improving underwater visual tracking with a large scale dataset and image enhancement. *arXiv preprint arXiv:2308.15816*, 2023.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip H. S. Torr. Staple: Complementary learners for real-time tracking. In *CVPR*, pp. 1401–1409. IEEE Computer Society, 2016.

Goutam Bhat, Joakim Johnander, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Unveiling the power of deep tracking. In *ECCV (2)*, volume 11206 of *Lecture Notes in Computer Science*, pp. 493–509. Springer, 2018.

Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Levi Cai, Nathan E McGuire, Roger Hanlon, T Aran Mooney, and Yogesh Girdhar. Semi-supervised visual tracking of marine animals using autonomous underwater vehicles. *International Journal of Computer Vision*, 131(6):1406–1427, 2023.

Wenrui Cai, Qingjie Liu, and Yunhong Wang. Hiptrack: Visual tracking with historical prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19258–19267, 2024.

Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19830–19843, 2023a.

Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *CVPR*, pp. 14572–14581. IEEE, 2023b.

Xin Chen, Ben Kang, Wanting Geng, Jiawen Zhu, Yi Liu, Dong Wang, and Huchuan Lu. Sutrack: Towards simple and unified single object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2239–2247, 2025.

Zhen Cheng, Fei Zhu, Xu-Yao Zhang, and Cheng-Lin Liu. Breaking the limits of reliable prediction via generated data. *International Journal of Computer Vision*, 133(3):1195–1221, 2025.

Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13608–13618, 2022.

Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: accurate tracking by overlap maximization. In *CVPR*, pp. 4660–4669. Computer Vision Foundation / IEEE, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5374–5383, 2019.

Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. Lasot: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision*, 129:439–461, 2021.

Qi Feng, Vitaly Ablavsky, Qinxun Bai, Guorong Li, and Stan Sclaroff. Real-time visual object tracking with natural language description. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 700–709, 2020.

Qi Feng, Vitaly Ablavsky, Qinxun Bai, and Stan Sclaroff. 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*, pp. 5851–5860, 2021.

Xiaokun Feng, Xuchen Li, Shiyu Hu, Dailing Zhang, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. Memvlt: Vision-language tracking with adaptive memory-based prompts. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Xiaokun Feng, Shiyu Hu, Xuchen Li, Dailing Zhang, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. Atctrack: Aligning target-context cues with dynamic target states for robust vision-language tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19850–19861, 2025a.

Xiaokun Feng, Dailing Zhang, Shiyu Hu, Xuchen Li, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. Enhancing vision-language tracking by effectively converting textual cues into visual cues. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

Shenyuan Gao, Chunluan Zhou, and Jun Zhang. Generalized relation modeling for transformer tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18686–18695, 2023.

Shi-Ran Ge, Jie Cao, and Ran He. Improving object detection models via llm-based training data synthesis. *International Journal of Computer Vision*, pp. 1–16, 2025.

Salma P González-Sabbagh and Antonio Robles-Kelly. A survey on underwater computer vision. *ACM Computing Surveys*, 55(13s):1–39, 2023.

Mingzhe Guo, Zhipeng Zhang, Heng Fan, and Liping Jing. Divert more attention to vision-language tracking. *Advances in Neural Information Processing Systems*, 35:4446–4460, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, et al. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19079–19091, 2024.

Shiyu Hu, Dailing Zhang, Xiaokun Feng, Xuchen Li, Xin Zhao, Kaiqi Huang, et al. 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, 2024.

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

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Kakani Katija, Paul LD Roberts, Joost Daniels, Alexandra Lapides, Kevin Barnard, Mike Risi, Ben Y Ranaan, Benjamin G Woodward, and Jonathan Takahashi. Visual tracking of deepwater animals using machine learning-controlled robotic underwater vehicles. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 860–869, 2021.

Xiaohai Li, Bineng Zhong, Qihua Liang, Zhiyi Mo, Jian Nong, and Shuxiang Song. Dynamic updates for language adaptation in visual-language tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19165–19174, 2025a.

Xin Li, Yuqing Huang, Zhenyu He, Yaowei Wang, Huchuan Lu, and Ming-Hsuan Yang. Citetracker: Correlating image and text for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9974–9983, 2023a.

Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. Diffusion models for image restoration and enhancement: a comprehensive survey. *International Journal of Computer Vision*, pp. 1–31, 2025b.

Yihao Li, Jun Yu, Zhongpeng Cai, and Yuwen Pan. Cross-modal target retrieval for tracking by natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4931–4940, 2022.

Yunfeng Li, Bo Wang, Ye Li, Zhuoyan Liu, Wei Huo, Yueming Li, and Jian Cao. Underwater object tracker: Uostrack for marine organism grasping of underwater vehicles. *Ocean Engineering*, 285: 115449, 2023b.

Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6495–6503, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pp. 11451–11461. IEEE, 2022.

Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial intelligence*, 293:103448, 2021.

Xingyu Luo, Yidong Cai, Jie Liu, Jie Tang, Gangshan Wu, and Limin Wang. Glad: Generative language-assisted visual tracking for low-semantic templates: X. luo et al. *International Journal of Computer Vision*, 134(3):121, 2026.

Yinchao Ma, Yuyang Tang, Wenfei Yang, Tianzhu Zhang, Jinpeng Zhang, and Mengxue Kang. Unifying visual and vision-language tracking via contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4107–4116, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pp. 11–20. IEEE Computer Society, 2016.

Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 300–317, 2018.

Karen Panetta, Landry Kezebou, Victor Oludare, and Sos Agaian. Comprehensive underwater object tracking benchmark dataset and underwater image enhancement with gan. *IEEE Journal of Oceanic Engineering*, 47(1):59–75, 2021.

Liang Peng, Junyuan Gao, Xinran Liu, Weihong Li, Shaohua Dong, Zhipeng Zhang, Heng Fan, and Libo Zhang. Vasttrack: Vast category visual object tracking. *arXiv preprint arXiv:2403.03493*, 2024.

Koutilya Pnvr, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, and David Jacobs. Ld-znet: A latent diffusion approach for text-based image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4157–4168, 2023.

Weize Quan, Jiaxi Chen, Yanli Liu, Dong-Ming Yan, and Peter Wonka. Deep learning-based image and video inpainting: A survey. *International Journal of Computer Vision*, 132(7):2367–2400, 2024.

Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11536–11546, 2023.

Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models. *arXiv preprint arXiv:2210.10606*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Yanyan Shao, Shuting He, Qi Ye, Yuchao Feng, Wenhan Luo, and Jiming Chen. 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*, pp. 19208–19217, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021.

Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson W. H. Lau, and Ming-Hsuan Yang. VITAL: visual tracking via adversarial learning. In *CVPR*, pp. 8990–8999. Computer Vision Foundation / IEEE Computer Society, 2018.

Yunjie Tian, Lingxi Xie, Jihao Qiu, Jianbin Jiao, Yaowei Wang, Qi Tian, and Qixiang Ye. Fast-itpn: Integrally pre-trained transformer pyramid network with token migration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13763–13773, 2021a.

Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip H. S. Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? In *NeurIPS*, pp. 726–738, 2021b.

Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9697–9706, 2023.

Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16293–16303, 2022.

You Wu, Xucheng Wang, Xiangyang Yang, Mengyuan Liu, Dan Zeng, Hengzhou Ye, and Shuiwang Li. Learning occlusion-robust vision transformers for real-time uav tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17103–17113, 2025.

Xue-Han Xu, Tao Jing, and Qing-Hao Meng. Usbtrack: Efficient underwater single-branch tracker via template prompt learning. *IEEE Sensors Journal*, 2025.

Zhengyuan Yang, Tushar Kumar, Tianlang Chen, Jingsong Su, and Jiebo Luo. Grounding-tracking-integration. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(9):3433–3443, 2020.

Dana R Yoerger, Annette F Govindarajan, Jonathan C Howland, Joel K Llopiz, Peter H Wiebe, Molly Curran, Justin Fujii, Daniel Gomez-Ibanez, Kakani Katija, Bruce H Robison, et al. A hybrid underwater robot for multidisciplinary investigation of the ocean twilight zone. *Science Robotics*, 6(55):eabe1901, 2021.

Chunhui Zhang, Guanjie Huang, Li Liu, Shan Huang, Yinan Yang, Xiang Wan, Shiming Ge, and Dacheng Tao. Webuav-3m: A benchmark for unveiling the power of million-scale deep uav tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9186–9205, 2022.

Chunhui Zhang, Li Liu, Guanjie Huang, Hao Wen, Xi Zhou, and Yanfeng Wang. Webuot-1m: Advancing deep underwater object tracking with a million-scale benchmark. *Advances in Neural Information Processing Systems*, 37:50152–50167, 2024a.

Chunhui Zhang, Li Liu, Guanjie Huang, Zhipeng Zhang, Hao Wen, Xi Zhou, Shiming Ge, and Yanfeng Wang. Underwater camouflaged object tracking meets vision-language sam2. *arXiv preprint arXiv:2409.16902*, 2024b.

Guangtong Zhang, Bineng Zhong, Qihua Liang, Zhiyi Mo, and Shuxiang Song. Diffusion mask-driven visual-language tracking. In *IJCAI*, pp. 1652–1660. ijcai.org, 2024c.

Qiuyang Zhang, Wei Song, Cong Liu, and Minghua Zhang. Putrack: Improved underwater object tracking via progressive prompting. *IEEE Transactions on Industrial Informatics*, 2025a.

Zhiguo Zhang, Zhiqing Guo, Liejun Wang, and Yongming Li. Ctiftrack: Continuous temporal information fusion for object track. *Expert Systems with Applications*, 262:125654, 2025b.

Bo Zhao, Weidong Yin, Lili Meng, and Leonid Sigal. Layout2image: Image generation from layout. *International journal of computer vision*, 128(10):2418–2435, 2020.

Haojie Zhao, Xiao Wang, Dong Wang, Huchuan Lu, and Xiang Ruan. Transformer vision-language tracking via proxy token guided cross-modal fusion. *Pattern Recognition Letters*, 168:10–16, 2023.

Yaozong Zheng, Bineng Zhong, Qihua Liang, Guorong Li, Rongrong Ji, and Xianxian Li. Toward unified token learning for vision-language tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2125–2135, 2023.

Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhiyi Mo, Shengping Zhang, and Xianxian Li. Odtrack: Online dense temporal token learning for visual tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 7588–7596, 2024.

Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23151–23160, 2023.

A Appendix

Recent studies Luo et al. (2026) have explored diffusion-based generative models for tracking by integrating visual templates and textual descriptions into the diffusion process, enabling linguistic semantics to complement target appearance. Building on this idea, they developed a generative Vision–Language Tracking framework that leverages the strong cross-modal interaction capability of generative models to facilitate tracking through the direct use of intermediate multi-modal representations, ultimately achieving strong tracking performance.

The diffusion-based tracking pipelines is commonly implemented using Stable Diffusion as the generative backbone. Built upon Latent Diffusion Models (LDMs) Song & Ermon (2019), which extend Denoising Diffusion Probabilistic Models (DDPMs) Ho et al. (2020); Sohl-Dickstein et al. (2015) into a compact latent space, Stable Diffusion reduces computational cost while preserving perceptually relevant information. Specifically, input images are encoded into latent variables via a VAE, and the diffusion model learns to predict added noise at each timestep using a time-conditioned U-Net optimized with the latent diffusion objective:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]. \tag{7}$$

where z_t is the input image x encoded into latent variables through a VAE-based encoder, t denotes the diffusion timestep uniformly sampled over the diffusion horizon and ϵ represents Gaussian noise.

This time-conditioned U-Net enables multi-scale feature modeling and cross-attention between visual and linguistic inputs, providing an efficient framework. Under this framework, diffusion-based tracking pipelines can extract generative features for tracking from the latent-space U-Net. Specially, a template image is first encoded via VAE and then perturbed through the forward diffusion process:

$$\begin{cases} \mathcal{F}_{latent} = \text{VAE}(template) \\ \mathcal{F}^0_{latent} = \text{Forward}(\mathcal{F}_{latent}) \end{cases}, \tag{8}$$

where $\mathcal{F}^0_{latent}$ denotes the noised latent features. The diffusion U-Net, conditioned on textual embeddings, progressively restores these features while enabling multi-modal interactions at different stages. Concretely, each Transformer-based submodule fuses latent features with text embeddings:

$$\begin{cases} \mathcal{F}_{text} = \text{TextEncoder}(text) \\ \mathcal{F}^s_u = \text{Transformer}_s(\mathcal{F}^{s-1}_{latent}, \mathcal{F}_{text}) \end{cases}, \quad s = 1, 2, ..., 16, \tag{9}$$

producing intermediate features $\mathcal{F}^s_u$ that capture progressively richer semantic abstractions. Prior works often reuse a subset of these features as language-enriched enhanced template representations for tracking.

In summary, generative-model-based tracking methods leverage latent-space diffusion models to enhance visual templates by extracting intermediate U-Net features conditioned on descriptions.

Hierarchical cross-modal fusion progressively injects language information into visual representations, producing semantically enriched template features without requiring fully reconstructed images. This approach exploits diffusion priors to provide more informative features, improving the tracker's ability to capture the target's appearance and semantics.