VL-UniTrack: A Unified Framework with Visual-Language Prompts for UAV-Ground Visual Tracking

Boyue Xu, *Student Member, IEEE*, Ruichao Hou, *Member, IEEE*, Tongwei Ren, *Member, IEEE* and Gangshan Wu, *Member, IEEE*

***Abstract*—UAV-ground visual tracking (UGVT) aims to simultaneously track the same target from both UAV and ground views. However, existing two-stream methods suffer from isolated feature extraction and rely heavily on implicit appearance matching, which often fails to establish reliable cross-view correspondence under drastic viewpoint discrepancies, thereby leading to unstable tracking. To address these issues, we propose VL-UniTrack, a unified framework enhanced by visual-language prompts. By encoding features from both views with a single shared encoder, our method reduces feature isolation between the two views and enables sufficient cross-view interaction. To alleviate the ambiguity caused by appearance-only matching, we further design a visual-language viewpoint prompting module that fuses language descriptions with visual features to generate learnable prompts. These prompts are then injected into a prompt-guided cross-view adapter to facilitate cross-view interaction and guide the learning of view-specific feature representations. In addition, we introduce a confidence-modulated mutual distillation loss to regularize training by suppressing noise propagation. Extensive experiments demonstrate that VL-UniTrack achieves state-of-the-art performance on the latest benchmark. The code is available at https://github.com/xuboyue1999/VL-UniTrack.git.**

***Index Terms*—UAV-ground visual tracking, multi-view tracking, visual-language prompts, unified framework, cross-view adapter.**

I. Introduction

UAV-GROUND visual tracking (UGVT) [1] aims to simultaneously track the same target from both UAV and ground views. By exploiting the complementary information from these two views, UGVT can alleviate the inherent limitations of single-view tracking, such as occlusion and restricted field of view, and therefore holds great potential for applications such as security surveillance and rescue missions [2], [3]. However, the substantial appearance discrepancy between UAV and ground views makes reliable cross-view feature interaction difficult to establish, rendering traditional tracking methods that rely primarily on visual similarity less effective.

To track the same target across different views, existing methods typically adopt a two-stream architecture, where independent visual encoders are used to extract features from the UAV and ground views separately, and feature interaction is performed only at a later stage, as illustrated in Fig. 1(a). Such methods are constrained by two limitations: (1) isolated feature extraction, where separate encoders prevent sufficient interaction between cross-view features; and (2) implicit appearance matching, where feature interaction

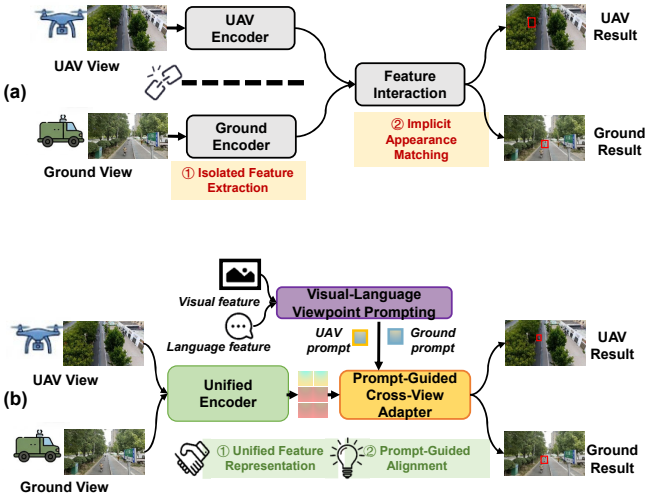


Fig. 1: Framework comparison between existing trackers and the proposed method (VL-UniTrack). (a) Existing tracker: Utilizes a split two-stream architecture with independent encoders, which relies on isolated feature extraction and appearance matching, leading to unreliable feature interaction under drastic viewpoint discrepancies. (b) VL-UniTrack: Uses a unified joint-encoding framework with a single shared encoder, incorporating visual-language viewpoint prompting for explicit view prompts and a prompt-guided cross-view adapter to guide viewpoint-aware interaction within the unified space.

This work was supported by the National Natural Science Foundation of China (92582103), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM118), the "111 Center" (No. B26023), the Research and Innovation Plan for Graduate Students in Jiangsu Province (No. 26CXJH0360), and the Collaborative Innovation Center of Novel Software Technology and Industrialization. *(Corresponding author: Tongwei Ren.)*

Boyue Xu and Gangshan Wu are with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210008, China (e-mail: xuby@smail.nju.edu.cn; gswu@nju.edu.cn).

Ruichao Hou is with the School of Artificial Intelligence, China Pharmaceutical University, Nanjing 211198, China (e-mail: rchou@cpu.edu.cn).

Tongwei Ren is with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210008, China, and also with the School of Artificial Intelligence, China Pharmaceutical University, Nanjing 211198, China (e-mail: rentw@nju.edu.cn).

© 2026 IEEE. Author accepted manuscript. Accepted for publication in *IEEE Transactions on Multimedia*.

relies solely on visual similarity and therefore becomes unreliable under drastic cross-view discrepancies, often failing to establish reliable correspondence. The UGVT task presents distinct requirements compared with those of one-stream single-view tracking and aligned multi-modal tracking (*e.g.*, visible-thermal tracking). In UGVT, the UAV and ground views are inherently unaligned and exhibit large viewpoint and appearance discrepancies, so methods developed for these settings may not fully address these cross-view characteristics when directly transferred.

To address these limitations, we propose VL-UniTrack, a unified framework for joint UAV-ground representation learning, as illustrated in Fig. 1(b). First, to address isolated feature extraction, we design a unified joint encoder. Unlike conventional two-stream methods, our framework takes UAV and ground views as joint inputs to a single visual encoder. This design reduces the separation between the two views, allowing cross-view features to be learned within a shared representation space and thereby enabling more thorough interaction. Second, to alleviate the unreliability of appearance-based matching, we incorporate visual-language prompts into the framework. Specifically, we encode view-specific cues as language prompts and fuse them with target appearance features to guide the learning of view-specific feature representations, thereby improving feature discrimination across views and bridging the large viewpoint gap. The unified encoder first establishes a shared cross-view representation through sufficient interaction between the two views, after which viewpoint-specific adaptation uses the corresponding prompts to emphasize the cues required by each tracking head.

Specifically, our method consists of three key modules. First, we introduce a visual-language viewpoint prompting (VLVP) module, which leverages CLIP [4] to encode view descriptions and fuse them with visual features, thereby generating learnable prompt tokens for the unified representation. Second, we design a prompt-guided cross-view adapter (PCVA), which injects these prompts into cross-view interaction to explicitly guide the learning of view-specific feature representations. Finally, to regularize the learning of the unified representation, we employ a confidence-modulated mutual distillation (CMD) loss, which dynamically prioritizes the more reliable view and suppresses noise propagation during training. Extensive experiments demonstrate that our method achieves state-of-the-art performance on the latest benchmark.

In summary, our contributions are as follows:

- We propose VL-UniTrack, a unified joint-encoding framework for unaligned UAV-ground visual tracking. By employing a single shared encoder for joint feature extraction, our method reduces the feature isolation caused by separate encoding and enables sufficient cross-view interaction.
- We introduce a visual-language viewpoint prompting module, which leverages a vision-language model to generate learnable prompt tokens. These prompts explicitly preserve view-specific cues within the unified representation.
- We design a prompt-guided cross-view adapter, which injects the generated prompts into cross-view interaction to guide the learning of more discriminative view-specific feature representations.
- We propose a confidence-modulated mutual distillation loss to regularize the learning of the unified representation by dynamically suppressing noise propagation from less reliable views.

II. Related Work

A. UAV-Ground Visual Tracking

UGVT effectively alleviates the limitations of single-view tracking and enhances overall perception capability. Early studies mainly relied on task allocation or late-stage fusion across different views for collaborative surveillance. For example, Minaeian *et al.* [5], [6] used UAVs for coarse-grained detection and then employed unmanned ground vehicles for high-resolution localization. These approaches typically perform perception independently in each view and combine the results through geometric transformations or rule-based integration.

Recent UGVT studies have increasingly explored multi-view representation learning and achieved notable progress. For instance, Zhou *et al.* [7] adopted graph neural networks to facilitate cross-view information fusion. Sun *et al.* [1] established a comprehensive UAV-ground tracking benchmark and proposed MVCL, which correlates and learns from feature cues extracted separately from each view. Similarly, Wang *et al.* [8] developed a collaborative perception framework in which features are extracted independently and then fused in a bird's-eye-view space using specialized collaboration modules. Hou *et al.* [9] proposed a multi-object tracking dataset for autonomous driving scenarios and provided solutions to occlusion challenges.

Although these methods have achieved significant progress, they still rely on separate feature extraction pipelines for different views and essentially perform tracking on each view individually. In contrast, our method adopts a unified framework for joint modeling and mutual supervision across both views. By leveraging unified token modeling together with visual-language-guided feature representation, the proposed method promotes more effective cross-view interaction while maintaining a compact overall framework.

B. Multi-View Visual Tracking

Current research on multi-view visual tracking mainly focuses on three categories: multi-UAV single-object tracking (SOT), multi-UAV multi-object tracking, and multi-camera object tracking.

In the area of multi-UAV SOT, Zhu *et al.* [10] established the MDOT benchmark and proposed ASNet, which employs independent Siamese trackers with template sharing. However, this framework still adopts separate view branches and relies on relatively weak interaction between views. As a result, the target is essentially tracked independently in each view, and the complementary information across views is not fully exploited.

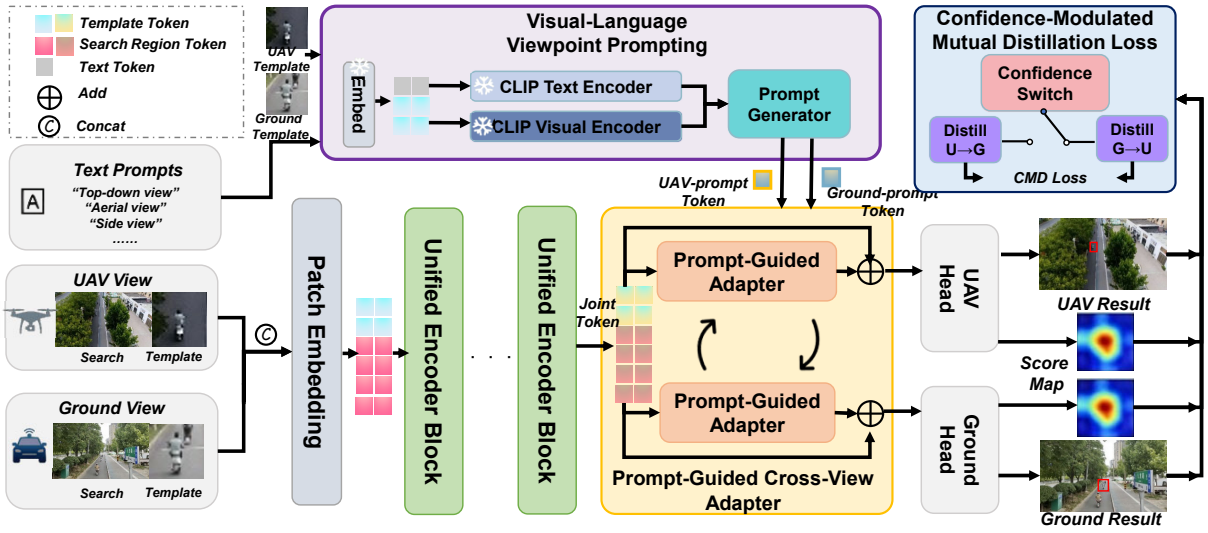


Fig. 2: The framework of the proposed VL-UniTrack method. Features of templates and search regions of UAV and ground view extracted by unified encoder are enhanced by the Prompt-Guided Cross-View Adapter. The PCVA is driven by viewpoint tokens generated via visual-language viewpoint prompting, which fuses text prompts with visual templates. The method is optimized using confidence-modulated mutual distillation loss, where a confidence switch determines the direction of distillation to mitigate error propagation from unreliable view.

In the multi-UAV tracking setting, Liu *et al.* [11] proposed MIA-Net, which uses homography-based geometric alignment to associate targets across different views. In addition, TransMDOT [12] and CRM [13] leverage self-attention and view-invariant modules to enhance feature interaction. Nevertheless, these methods still follow a pipeline in which features are first extracted separately and then aligned or interacted afterward.

Similarly, multi-camera object tracking also suffers from isolated cross-view feature modeling. For example, Divotrack [14] extends single-view trackers to the multi-view setting, but still processes each view independently. GMT [15] employs Transformers to model cross-view appearance and spatio-temporal features for similarity-based matching. More recently, MITracker [16] performs multi-view bounding-box tracking through geometric projection and feature fusion, providing another approach to information integration across views. In general, these methods rely on independent view branches, where features are processed in parallel pipelines before later interaction. This design limits the full exploitation of complementary information across views.

In summary, most existing multi-view tracking frameworks rely on independent encoders and are characterized by weak interaction or late-stage fusion. As a result, they fail to achieve sufficiently deep cross-view interaction and lack explicit guidance for unified joint modeling.

C. Single-View visual Tracking

Single-view visual tracking can generally be categorized into classification-based, Siamese-based, and Transformer-based methods. Classification-based trackers, such as MDNet [17] and DiMP [18], formulate tracking as an online foreground-background discrimination problem. Although these methods maintain strong robustness through continuous

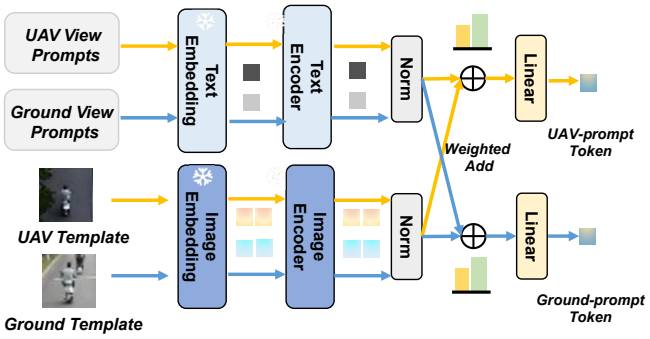


Fig. 3: Detailed design of the visual-language viewpoint prompting module, which generates view-specific prompt tokens by fusing textual viewpoint priors with template visual features.

model updates, they often suffer from high computational cost. To improve efficiency, Siamese-based trackers, including SiamFC [19] and LightTrack [20], adopt a dual-branch framework to match features between templates and search regions. While these methods achieve real-time performance, they rely heavily on static templates, which limits their adaptability to significant appearance changes and long-term tracking scenarios.

More recently, Transformer-based methods have been introduced to enable stronger feature modeling. Representative approaches, such as OSTrack [21] and SwinTrack [22], replace conventional encoders with Transformer architectures, enabling more comprehensive feature extraction and relation modeling. Furthermore, sequential trackers such as SeqTrack [23] and ARTrack [24] incorporate temporal cues to improve localization accuracy. Nevertheless, extending

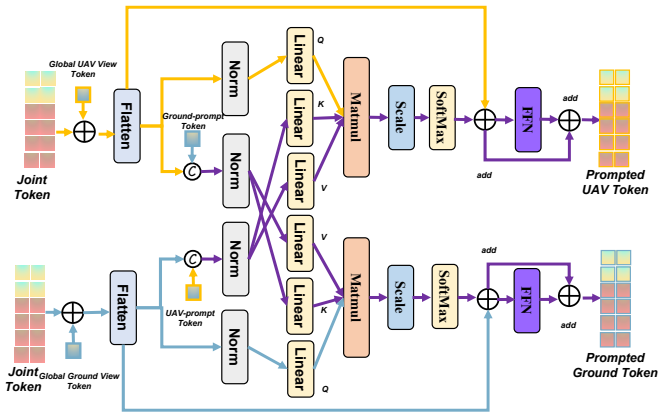


Fig. 4: Detailed design of the prompt-guided cross-view adapter, which injects viewpoint prompts into cross-view interaction to learn more discriminative view-specific representations.

such single-view methods to multi-view tracking remains challenging, as it requires handling the complex interaction and heterogeneity across different views.

III. Method

A. Overview

The overall framework of VL-UniTrack is illustrated in Fig. 2. Given the UAV template-search pair $\mathcal{I}_u = \{Z_u, X_u\}$ and the ground-view pair $\mathcal{I}_g = \{Z_g, X_g\}$, we first form paired target-centered inputs by channel-wise concatenation. The unified visual encoder jointly projects the paired patches and produces a single joint feature representation F_{J}. The VLVP module employs a frozen CLIP encoder to generate the view-specific prompt tokens P_u and P_g from textual viewpoint descriptions and visual templates. Conditioned on these prompts, PCVA transforms F_{J} into the enhanced UAV- and ground-specific representations $\hat{F}_u$ and $\hat{F}_g$. Parallel prediction heads then output the score maps and bounding boxes for the two views. During training, CMD applies confidence-gap-gated directional supervision between the two enhanced view-specific representations.

B. Unified Visual Encoder

Traditional UAV-ground visual tracking methods typically adopt a dual-stream architecture with separate encoders for UAV and ground views. Such independent feature extraction suffers from a semantic gap between the two views, which hinders effective cross-view interaction. To address this limitation, we employ a unified visual encoder [25] to project UAV and ground-view features into a shared semantic space. This design facilitates more thorough cross-view interaction and provides a solid foundation for subsequent learning of view-specific representations.

Given target-centered UAV and ground-view template crops Z_u and Z_g, and their search-region crops X_u and X_g, we resize the crops in each pair to the same spatial resolution and form $\bar{Z} = [Z_u; Z_g]_c$ and $\bar{X} = [X_u; X_g]_c$, where $[\cdot;\cdot]_c$ denotes channel-wise concatenation. Each spatial patch of $\bar{Z}$ or $\bar{X}$ therefore contains six channels formed from the two target-centered views at the same normalized index.

The six-channel joint projection first fuses local patches at the same spatial index in the two views to produce intermediate features:

$$\begin{aligned} \mathbf{h}_k^z &= \mathbf{W}_6[\mathbf{z}_{u,k}; \mathbf{z}_{g,k}] + \mathbf{b}, \\ \mathbf{h}_k^x &= \mathbf{W}_6[\mathbf{x}_{u,k}; \mathbf{x}_{g,k}] + \mathbf{b}, \end{aligned} \tag{1}$$

where $\mathbf{z}_{u,k}, \mathbf{z}_{g,k}$ and $\mathbf{x}_{u,k}, \mathbf{x}_{g,k}$ are flattened three-channel template and search patches at position k, respectively; p is the initial patch size; $\mathbf{W}_6 \in \mathbb{R}^{D_0 \times 6p^2}$ and $\mathbf{b} \in \mathbb{R}^{D_0}$ are the joint projection weights and bias; and $\mathbf{h}_k^z, \mathbf{h}_k^x \in \mathbb{R}^{D_0}$ are intermediate features, with D_0 denoting the initial projection's output channel dimension. The superscripts z and x distinguish the template and search regions, respectively.

Let $\mathcal{P}_6(\cdot)$ denote joint tokenization, comprising the six-channel projection, the existing hierarchical front-end processing of Fast-iTPN, and spatial unfolding. Its outputs satisfy $\mathcal{P}_6(\bar{Z}) \in \mathbb{R}^{N_z \times D}$ and $\mathcal{P}_6(\bar{X}) \in \mathbb{R}^{N_x \times D}$, where N_z, N_x are the template and search token counts after front-end processing, and D is the dimension of each token, with one token per row. We concatenate the search and template tokens along the sequence dimension and add positional embeddings:

$$E_{\mathrm{J}} = \left[\mathcal{P}_6(\bar{X})\,;\, \mathcal{P}_6(\bar{Z})\right] + \Pi \in \mathbb{R}^{(N_x+N_z)\times D}, \tag{2}$$

where E_{J} is the joint input sequence, the upright subscript J denotes a joint representation, $\Pi \in \mathbb{R}^{(N_x+N_z)\times D}$ is the positional embedding matrix, and $[\cdot;\cdot]$ denotes concatenation along the token dimension. Joint tokenization couples information from the two views within the tokens. The search and template tokens form a single joint sequence of length $N_x + N_z$, which is passed to the subsequent encoding stages:

$$F_{\mathrm{J}} = \Phi(E_{\mathrm{J}}), \tag{3}$$

where $\Phi(\cdot)$ denotes the subsequent Fast-iTPN encoding and search-feature readout. The readout extracts the N_x search-region tokens from the encoded sequence and reshapes them into the joint feature map $F_{\mathrm{J}} \in \mathbb{R}^{D\times H\times W}$, with spatial dimensions H, W satisfying $N_x = HW$. PCVA then adapts this joint feature map into view-specific representations.

C. Visual-Language Viewpoint Prompting

The unified encoder produces a compact joint representation in which information from the paired UAV and ground-view inputs has already been coupled. Although unified encoding reduces feature isolation, viewpoint-specific cues may still be insufficiently represented within the unified representation. We therefore introduce VLVP to provide explicit UAV and ground-view semantic guidance for deriving view-conditioned representations from F_{J}. By leveraging the knowledge of CLIP [4], VLVP generates semantic tokens that serve as viewpoint prompts and preserve the viewpoint-specific characteristics required for robust tracking.

The visual branch in VLVP extracts target-specific appearance context from the initial target templates, while the language branch provides generic semantics for the UAV and ground viewpoints. Their fusion produces prompts that encode both target appearance and viewpoint information to

guide cross-view interaction and view-specific representation learning. In our tracking setting, the target is specified by its initial bounding boxes and visual templates, and language descriptions provide viewpoint guidance.

As illustrated in Fig. 3, VLVP has parallel text and vision branches. To capture view-specific cues, the text branch averages CLIP embeddings of viewpoint descriptions:

$$P_v^{\mathrm{L}} = \frac{1}{|\mathcal{T}_v|} \sum_{t \in \mathcal{T}_v} \Phi_{\mathrm{L}}(t), \quad v \in \{u, g\}, \tag{4}$$

where $\Phi_{\mathrm{L}}(\cdot)$ denotes the language branch implemented using the frozen CLIP text encoder, $\mathcal{T}_v$ is the set of descriptions for view v, $t \in \mathcal{T}_v$ is an individual description, and P_v^{L} is the language prompt feature before fusion. The UAV descriptions are "Top-down view", "Aerial view" and "Overhead shot", while the ground-view descriptions are "Side view", "Front view" and "Street level view". The vision branch encodes the initial templates to extract target appearance information:

$$P_v^{\mathrm{V}} = \Phi_{\mathrm{V}}(Z_v), \quad v \in \{u, g\}, \tag{5}$$

where $\Phi_{\mathrm{V}}(\cdot)$ denotes the visual branch implemented using the frozen CLIP image encoder, Z_v is the initial template image for view v, and P_v^{V} is the visual prompt feature before fusion. Weighted fusion and linear projection produce the final prompt tokens:

$$P_v = \varphi\left(\mathrm{Norm}(P_v^{\mathrm{L}}) + \alpha \cdot \mathrm{Norm}(P_v^{\mathrm{V}})\right), \quad v \in \{u, g\}, \tag{6}$$

where $\mathrm{Norm}(\cdot)$ denotes L2 normalization, α is a learnable balancing parameter, and $\varphi(\cdot)$ projects features linearly into the tracker's embedding dimension D. The tokens P_u and P_g jointly encode viewpoint semantics and visual context to guide subsequent cross-view interaction. The features $P_v^{\mathrm{L}}, P_v^{\mathrm{V}}$ are the pre-fusion prompt features within VLVP, while P_v is the final fused prompt passed to PCVA. Unlike a static view-identity embedding that only indicates whether a branch corresponds to the UAV or ground view, the proposed prompts jointly encode high-level viewpoint semantics and target-specific visual context extracted from the initial templates. They therefore provide instance-conditioned guidance for subsequent cross-view interaction.

D. Prompt-Guided Cross-View Adapter

To effectively exploit the complementary information between UAV and ground views, we propose a PCVA. This module is designed to establish viewpoint-aware interaction across the two views. Specifically, the viewpoint prompts generated by the VLVP module are incorporated into a cross-attention mechanism, so that cross-view interaction is guided by view-specific cues rather than relying solely on visual similarity. In PCVA, Cross-attention establishes the basic information exchange between the two views, the global view tokens provide branch identity cues, and the VLVP prompts introduce explicit viewpoint-aware visual-language guidance into the interaction.

Fig. 4 illustrates the internal prompt-guided interaction of PCVA. The module takes the joint feature tensor F_{J} as input and processes it through two symmetric branches. First, to adapt the unified features to each specific view, we inject learnable global view tokens through element-wise addition:

$$F_u = \Psi(F_{\mathrm{J}} + \mathcal{G}_u), \tag{7}$$

$$F_g = \Psi(F_{\mathrm{J}} + \mathcal{G}_g), \tag{8}$$

where $\mathcal{G}_u, \mathcal{G}_g$ are learnable global view tokens; $\Psi(\cdot)$ denotes spatial unfolding, converting a $D \times H \times W$ feature map into an $HW \times D$ token matrix in spatial row-major order, with each row representing the D-dimensional feature at one spatial location; and F_u, F_g are the UAV and ground base features, respectively.

These base features are then refined through cross-attention-based interaction. Taking the UAV branch as an example, the goal is to enhance the UAV representation by querying complementary information from the ground view. To impose viewpoint guidance, we concatenate the VLVP-generated UAV prompt with the ground base features:

$$K_u = V_u = [P_u; F_g], \tag{9}$$

where $[\cdot;\cdot]$ denotes concatenation along the token dimension, P_u is the fused UAV prompt, F_g is the ground base feature, and K_u, V_u are the key/value projection inputs. The concatenated sequence is layer-normalized after prompt insertion, while the query is normalized separately. We then apply the attention and feed-forward residual sublayers and integrate the adapter output through learnable residual scaling:

$$\hat{F}_u = \gamma_u \,\mathrm{FFN}(\mathrm{MHCA}(F_u, K_u, V_u)) + F_u, \tag{10}$$

where the three inputs of $\mathrm{MHCA}(\cdot,\cdot,\cdot)$ correspond to query, key, and value, respectively; $\mathrm{MHCA}(\cdot,\cdot,\cdot)$ and $\mathrm{FFN}(\cdot)$ are pre-normalized multi-head cross-attention and feed-forward residual sublayers with the internal skip connections in Fig. 4; γ_u, γ_g are branch-specific learnable parameters; and $\hat{F}_u$ is the enhanced UAV feature. Symmetrically, the ground branch queries with F_g and constructs key/value inputs by concatenating P_g with F_u to obtain the enhanced ground feature $\hat{F}_g$.

E. Confidence-Modulated Mutual Distillation Loss

CMD lets the relatively more reliable view guide the other without forcing teacher assignment when their reliability is indistinguishable. It uses the maximum target-center response as an online relative reliability proxy:

$$c_v = \max_{i,j} \mathbf{R}_v(i,j), \qquad v \in \{u, g\}, \tag{11}$$

where $\mathbf{R}_v \in [0,1]^{H \times W}$ is the view-v probability response map, H, W are its spatial dimensions, (i, j) indexes locations, and c_v is the maximum response. This proxy represents the response strength at the most likely target location and requires no additional quality-estimation network. Importantly, CMD uses it only to compare the relative reliability of the two views rather than interpreting an individual response as an absolute probability of correct tracking.

During training, we perform target-region-weighted pooling on the view-specific PCVA features using the ground-truth Gaussian center heatmaps:

$$\mathbf{f}_v = \frac{\sum_{i,j} \mathbf{M}_v(i,j)\hat{F}_v(i,j)}{\sum_{i,j} \mathbf{M}_v(i,j) + \epsilon}, \qquad v \in \{u, g\}, \tag{12}$$

where $\mathbf{M}_v$ is the Gaussian center heatmap, $\hat{F}_v(i,j)$ the enhanced feature vector at the corresponding spatial location, $\mathbf{f}_v$ the pooled target feature, and ϵ a small positive constant for numerical stability. The pooled feature allows CMD to regularize target-relevant representations instead of directly matching the complete spatial feature maps.

We activate directional distillation only when the confidence gap exceeds a margin. The CMD objective is defined as

$$\begin{aligned}\mathcal{L}_{\text{cmd}} = \mathbb{I}(c_u > c_g + \delta)\mathcal{L}_{\cos}(\mathbf{f}_g, \text{sg}[\mathbf{f}_u]) \\ + \mathbb{I}(c_g > c_u + \delta)\mathcal{L}_{\cos}(\mathbf{f}_u, \text{sg}[\mathbf{f}_g]),\end{aligned} \tag{13}$$

where $\mathbb{I}(\cdot)$ is the indicator function, equal to 1 when the condition holds and 0 otherwise, $\text{sg}[\cdot]$ denotes stop-gradient, and $\delta = 0.05$ is the confidence-gap margin. The cosine loss is defined as

$$\mathcal{L}_{\cos}(\mathbf{r}_1, \mathbf{r}_2) = 1 - \frac{\mathbf{r}_1^\top \mathbf{r}_2}{\|\mathbf{r}_1\|_2 \|\mathbf{r}_2\|_2}, \tag{14}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the input feature vectors and $\|\cdot\|_2$ denotes the Euclidean norm. Thus, the higher-confidence view serves as a fixed teacher only when the relative difference is sufficiently clear. When $|c_u - c_g| \leq \delta$, neither term is activated and $\mathcal{L}_{\text{cmd}} = 0$. This confidence-gap gating filters samples with an ambiguous teacher direction, although it does not assume that every activated teacher is necessarily correct.

Algorithm 1 Confidence-Modulated Mutual Distillation (CMD) loss

Require: Enhanced features $\hat{F}_u, \hat{F}_g$; probability maps $\mathbf{R}_u, \mathbf{R}_g$; Gaussian maps $\mathbf{M}_u, \mathbf{M}_g$; margin δ
Ensure: Distillation loss $\mathcal{L}_{\text{cmd}}$

$c_u \leftarrow \max(\mathbf{R}_u)$; $c_g \leftarrow \max(\mathbf{R}_g)$
$\mathbf{f}_u \leftarrow \text{WeightedPool}(\hat{F}_u, \mathbf{M}_u)$
$\mathbf{f}_g \leftarrow \text{WeightedPool}(\hat{F}_g, \mathbf{M}_g)$
$\mathcal{L}_{\text{cmd}} \leftarrow 0$
if $c_u > c_g + \delta$ **then**
    $\mathcal{L}_{\text{cmd}} \leftarrow \mathcal{L}_{\cos}(\mathbf{f}_g, \text{sg}[\mathbf{f}_u])$
else if $c_g > c_u + \delta$ **then**
    $\mathcal{L}_{\text{cmd}} \leftarrow \mathcal{L}_{\cos}(\mathbf{f}_u, \text{sg}[\mathbf{f}_g])$
end if
return $\mathcal{L}_{\text{cmd}}$

Following existing center-based trackers [21], each view-specific prediction head contains a target-center score-map branch and a bounding-box regression branch. The score-map branch is supervised by a focal loss computed against a Gaussian center heatmap, while the predicted bounding box is optimized using L1 and generalized IoU losses. The overall objective further incorporates the proposed CMD loss:

$$\begin{aligned}\mathcal{L} = \sum_{v \in \{u,g\}} \left(\lambda_{\text{focal}}\mathcal{L}^v_{\text{focal}} + \lambda_1 \mathcal{L}^v_1 + \lambda_{\text{giou}}\mathcal{L}^v_{\text{giou}}\right) \\ + \lambda_{\text{cmd}}\mathcal{L}_{\text{cmd}},\end{aligned} \tag{15}$$

where $\mathcal{L}$ is the total loss, $\mathcal{L}^v_{\text{focal}}$, $\mathcal{L}^v_1$ and $\mathcal{L}^v_{\text{giou}}$ are the focal, L1, and GIoU losses for view v, respectively, and λ_{focal}, λ_1, λ_{giou} and λ_{cmd} are loss weights set to 1, 5, 2 and 1, respectively, in all experiments.

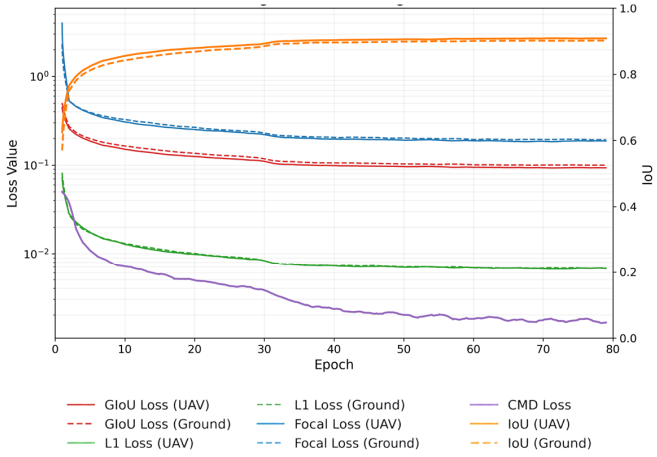


Fig. 5: Visualization results of Optimization Trajectory.

IV. Experiments

A. Dataset and Evaluation

We evaluate the proposed method on the UGVT dataset [1], which is currently the largest benchmark for UAV-ground visual tracking. It contains over 200 pairs of synchronized UAV and ground-view tracking sequences, covering diverse target categories such as pedestrians and vehicles. Moreover, the dataset includes a wide range of challenging scenarios, including severe occlusion, similar objects, camera motion, and large scale variation. The main challenge of UGVT lies in the drastic viewpoint discrepancy between UAV and ground views, which leads to significant appearance divergence for the same target and imposes stringent requirements on the tracker's ability to maintain robust cross-view consistency.

Following prior work, we adopt two evaluation metrics: Precision Rate (PR) and Success Rate (SR).

PR measures center-location accuracy at the 20-pixel threshold and is averaged over all test sequences:

$$\text{PR} = \frac{1}{|\mathcal{Q}|}\sum_{q \in \mathcal{Q}} \frac{1}{N_q}\sum_{i=1}^{N_q} \mathbb{I}(\text{CLE}_{q,i} \leq 20), \tag{16}$$

where $\mathcal{Q}$ is the test-sequence set, N_q is the frame count of sequence q, and $\text{CLE}_{q,i}$ is the Euclidean distance between predicted and ground-truth centers in its i-th frame, in pixels; $\mathbb{I}(\cdot)$ is the indicator function defined above.

For the q-th sequence, the success rate at IoU threshold τ is $S_q(\tau) = \frac{1}{N_q}\sum_{i=1}^{N_q} \mathbb{I}(\text{IoU}_{q,i} > \tau)$. SR is reported as the area



Fig. 6: Qualitative comparison between the proposed method and state-of-the-art methods on the UGVT test set. The ground truth is marked in green.

under the success curve, implemented as the mean success rate over 21 IoU thresholds:

$$\mathrm{SR} = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}}\frac{1}{|\Theta|}\sum_{\tau\in\Theta} S_q(\tau), \tag{17}$$

where $\Theta = \{0, 0.05, \ldots, 1\}$ denotes the set of IoU thresholds, $\mathrm{IoU}_{q,i}$ is the intersection over union between the predicted and ground-truth boxes in frame i of sequence q, and $S_q(\tau)$ is the success rate of that sequence at threshold τ; $\mathcal{Q}$ and N_q have the same meanings as in Eq. (16).

B. Implementation Details

The proposed method is trained on a server equipped with a CPU clocked at 5.2 GHz and a GPU with 32 GB of memory. The model is trained for 80 epochs with a batch size of 8, and each epoch contains 10,000 samples. We use AdamW with a base learning rate of 1×10^{-4} and an encoder learning rate of 1×10^{-5}, corresponding to an encoder multiplier of 0.1. The learning rate is decayed by a factor of 0.1 every 30 epochs, and the weight decay is set to 1×10^{-4}. All other training parameters are consistent with the baseline [26]. The model has 98.67M trainable parameters and achieves real-time performance at approximately 30 FPS.

C. Comparison with State-of-the-Arts Method

To comprehensively evaluate the effectiveness of the proposed method, we compare VL-UniTrack with a broad range of state-of-the-art trackers on the UGVT test set, including DiMP50 [18], SiamFC++ [27], TransT [28], KeepTrack [29], SparseTT [30], MixFormer [31], ToMP50 [32], OSTrack [21], AiATrack [33], CTTrack [34], MVCL [1], TransMDOT [12], CRM [13], MITracker [16], UNTrack [35], XTrack [36], SSTrack [37], and SuTrack [26]. The quantitative results are summarized in Table I. As shown in the table, VL-UniTrack

TABLE I: Comparison between the proposed method and state-of-the-art trackers on UGVT test set. The best results are highlighted in **bold**. The performance is evaluated in terms of PR and SR. * denotes multi-view tracking methods.

Method	Year	Ground View		UAV View		Average		FPS↑
		SR↑	PR↑	SR↑	PR↑	SR↑	PR↑	
DiMP50	ICCV19	0.567	0.815	0.606	0.793	0.587	0.804	35
SiamFC++	AAAI20	0.466	0.675	0.534	0.703	0.500	0.689	57
TransT	CVPR21	0.578	0.773	0.572	0.728	0.575	0.751	43
KeepTrack	ICCV21	0.599	0.818	0.636	0.821	0.618	0.820	16
SparseTT	IJCAI22	0.622	0.813	0.585	0.716	0.604	0.765	36
MixFormer	CVPR22	0.536	0.689	0.592	0.772	0.564	0.731	28
ToMP50	CVPR22	0.621	0.837	0.645	0.835	0.633	0.836	24
OSTrack	ECCV22	0.580	0.758	0.572	0.740	0.576	0.749	**58**
AiATrack	ECCV22	0.590	0.781	0.629	0.822	0.610	0.802	40
CTTrack	AAAI23	0.601	0.812	0.612	0.819	0.607	0.816	19
MVCL*	TCSVT23	0.637	0.844	0.623	0.786	0.630	0.815	21
TransMDOT*	TCSVT23	0.587	0.733	0.687	0.853	0.637	0.793	30
UNTrack	CVPR24	0.654	0.797	0.709	0.858	0.682	0.828	15
CRM*	TCSVT24	0.408	0.586	0.623	0.840	0.515	0.713	38
XTrack	ICCV25	0.666	0.851	0.700	0.868	0.683	0.859	17
SSTrack	AAAI25	0.612	0.766	0.702	0.874	0.657	0.820	33
SuTrack	AAAI25	0.587	0.745	0.649	0.818	0.618	0.781	40
MITracker*	CVPR25	0.549	0.694	0.641	0.798	0.595	0.746	16
Ours		**0.704**	**0.853**	**0.717**	**0.886**	**0.711**	**0.869**	30

TABLE II: Component analysis on UGVT test set. PCVA represents prompt-guided cross-view adapter, CMD represents confidence-modulated mutual distillation loss, and VLVP represents visual-language viewpoint prompting. $\Delta\,(\%)$ denotes the improvement over the baseline in the first row. The best results are highlighted in **bold**.

PCVA	CMD	VLVP	UAV View				Ground View			
			PR↑	Δ(%)	SR↑	Δ(%)	PR↑	Δ(%)	SR↑	Δ(%)
			0.818	–	0.649	–	0.745	–	0.587	–
✓			0.842	+2.4	0.681	+3.2	0.829	+8.4	0.684	+9.7
	✓		0.841	+2.3	0.683	+3.4	0.820	+7.5	0.679	+9.2
✓	✓		0.859	+4.1	0.700	+5.1	0.833	+8.8	0.690	+10.3
✓	✓	✓	**0.886**	+6.8	**0.717**	+6.8	**0.853**	+10.8	**0.704**	+11.7

TABLE III: Ablation study of VLVP module on UGVT test set. The best results are highlighted in **bold**. $\Delta\,(\%)$ denotes the improvement over w/o VLVP.

Setting	UAV View				Ground View			
	PR↑	Δ(%)	SR↑	Δ(%)	PR↑	Δ(%)	SR↑	Δ(%)
w/o VLVP	0.859	–	0.700	–	0.833	–	0.690	–
Learnable Embedding	0.869	+1.0	0.706	+0.6	0.846	+1.3	0.699	+0.9
Only Visual Prompt	0.878	+1.9	0.712	+1.2	0.850	+1.7	0.699	+0.9
Only Text Prompt	0.869	+1.0	0.705	+0.5	0.841	+0.8	0.695	+0.5
Ours	**0.886**	+2.7	**0.717**	+1.7	**0.853**	+2.0	**0.704**	+1.4

achieves the best performance on both the ground and UAV views, as well as on the overall average metrics, while still maintaining real-time inference at 30 FPS. These results demonstrate that the proposed unified framework can effectively exploit complementary information across the two views without sacrificing practical efficiency.

More specifically, on the ground view, VL-UniTrack outperforms the strongest competing method, XTrack, improving PR from 0.851 to 0.853 and SR from 0.666 to 0.704, corresponding to gains of 0.2% and 3.8%, respectively. On the UAV view, our method improves the best competing SR result from 0.709 to 0.717, and the best competing PR result from 0.874 to 0.886, yielding gains of 0.8% and 1.2%, respectively. In terms of the overall average performance, VL-UniTrack achieves gains over XTrack of 1.0% in PR and 2.8% in SR. These consistent improvements across both views indicate that our method does not favor one branch at the expense of the other; instead, it enables UAV and ground-view features to reinforce each other within a unified representation space, leading to more stable and reliable tracking under challenging cross-view conditions.

To further strengthen the experimental comparison directly related to UAV-ground tracking, we supplement the existing MVCL baseline with three representative multi-view trackers: TransMDOT, CRM, and MITracker. As shown in Table I, VL-UniTrack achieves the highest PR and SR on both views among these multi-view baselines. Compared with TransMDOT, which performs best on average among the three added methods, VL-UniTrack improves the average PR from 0.793 to 0.869 and the average SR from 0.637 to 0.711, corresponding to gains of 7.6% and 7.4%, respectively. These results further demonstrate the effectiveness of our method for collaborative UAV-ground tracking.

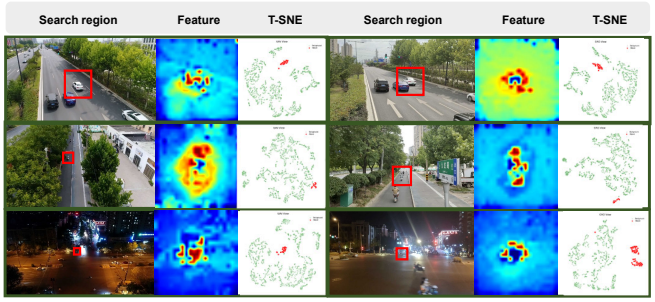


Fig. 7: Feature visualization and t-SNE results of the proposed method.

D. Ablation Study

To thoroughly validate the contribution of each component in the proposed framework, we conduct a series of component-wise ablation experiments, and the results are summarized in Table II.

The first row reports the baseline performance: 0.818 and 0.649 in UAV PR and SR, respectively, and 0.745 and 0.587 in ground-view PR and SR, respectively. After introducing the PCVA module, the model obtains gains of 2.4% and 3.2% on the UAV view, together with larger improvements of 8.4% and 9.7% on the ground view. This indicates that prompt-guided cross-view interaction is particularly beneficial when the target is more difficult to localize. Adding CMD alone on top of the baseline also consistently improves performance, yielding gains of 2.3% and 3.4% on the UAV view, as well as 7.5% and 9.2% on the ground view. These results verify that confidence-aware mutual distillation can effectively enhance cross-view collaboration during training.

When PCVA and CMD are used together, the performance is further improved, with gains of 4.1% and 5.1% on the UAV view and 8.8% and 10.3% on the ground view, indicating that the two modules are complementary. Finally, after incorporating VLVP, the full model achieves the best performance, with overall improvements of 6.8% and 6.8% on the UAV view and 10.8% and 11.7% on the ground view over the baseline. These results clearly demonstrate that each module contributes positively to the final performance, and that their combination yields the best performance among the evaluated configurations.

E. Targeted Analysis of Prompt Guidance in PCVA

To further isolate the roles of different components in PCVA, we progressively introduce Cross-attention, the global view tokens, and the VLVP prompts. As shown in Table IV, introducing Cross-attention without prompt information improves PR from 0.841 to 0.853 and SR from 0.683 to 0.698 on the UAV view, and PR from 0.820 to 0.832 and SR from 0.679 to 0.692 on the ground view. The resulting average PR and SR are 0.843 and 0.695, respectively. This result shows that basic cross-view attention can effectively exploit complementary information between the two views. Adding the global view tokens yields 0.846 and 0.695 in average PR and SR, respectively,

TABLE IV: Targeted ablation of prompt guidance in PCVA on the UGVT test set. The best results are in **bold**.

Setting	UAV View		Ground View		Average	
	PR↑	SR↑	PR↑	SR↑	PR↑	SR↑
CMD only / w/o PCVA	0.841	0.683	0.820	0.679	0.830	0.681
+ Cross-attention	0.853	0.698	0.832	0.692	0.843	0.695
+ Global view token	0.859	0.700	0.833	0.690	0.846	0.695
+ VLVP prompt	**0.886**	**0.717**	**0.853**	**0.704**	**0.869**	**0.711**

TABLE V: Comparison of encoder architectures on UGVT. Online FLOPs and latency are model-forward costs; the best results are in **bold**.

Architecture	Average		Efficiency	
	PR↑	SR↑	FLOPs↓	Latency↓
Independent Two-stream	0.825	0.684	410.6G	14.48 ms
Shared-weight Two-stream	0.838	0.691	410.6G	14.46 ms
Unified Joint Encoder (Ours)	**0.869**	**0.711**	**213.8**G	**7.98** ms

corresponding to comparatively small further changes. After incorporating the VLVP prompts, the model achieves the best average PR and SR of 0.869 and 0.711, respectively. Compared with Cross-attention without prompt information, the complete prompt-guided design achieves improvements in average PR and SR of 2.6% and 1.6%, respectively. These results show that Cross-attention establishes basic cross-view information exchange, the global view tokens provide branch identity cues, and the VLVP prompts further provide effective viewpoint-aware visual-language guidance for the interaction.

F. Encoder Architecture Analysis

To distinguish the effect of parameter sharing from unified representation learning, we compare three encoder architectures. Independent Two-stream employs two encoders with independent parameters and processes the UAV and ground views separately. Shared-weight Two-stream uses the same encoder weights for both views but still performs two separate forward passes, and therefore does not construct joint tokens or joint features inside the encoder. Our Unified Joint Encoder concatenates the paired target-centered inputs channel-wise, projects every paired patch into one view-coupled token through a six-channel joint patch embedding, and encodes the resulting compact joint sequence in a single Fast-iTPN forward pass. All three architectures retain the same VLVP, PCVA, CMD, and prediction heads, so the comparison focuses on the encoder design.

As shown in Table V, Independent Two-stream obtains 0.825 and 0.684 in average PR and SR, respectively. Sharing the encoder weights improves the average PR and SR to 0.838 and 0.691, respectively, showing that parameter sharing itself provides a certain benefit. However, because the two views are still encoded in separate forward passes, the online computation and latency remain almost unchanged. More importantly, our Unified Joint Encoder achieves 0.869 and 0.711 in average PR and SR, respectively, corresponding to gains of 3.1% and 2.0% over Shared-weight Two-stream. It also reduces Online FLOPs from 410.6G to 213.8G and Online Latency from 14.46 ms to 7.98 ms. These results show that the benefit is not attributable to parameter sharing alone, but also arises from the compact joint representation formed by paired-view joint patch projection. Compared with Independent Two-stream, our method achieves improvements in average PR and SR of 4.4% and 2.7%, respectively, while requiring substantially lower online computation, thereby achieving a better accuracy-efficiency balance.

TABLE VI: Comparison of score-map confidence measures for CMD teacher selection at matched activation coverage.

Confidence measure	Switch Acc.	Meaningful-gap Acc.
Normalized entropy	49.26%	52.30%
APCE	64.31%	72.70%
Maximum response (Ours)	**65.98**%	**74.47**%

G. Confidence Reliability Analysis

To examine whether the maximum score-map response used by CMD provides a reliable cue for teacher selection, we compare it with normalized entropy and average peak-to-correlation energy (APCE). Each measure is evaluated by whether it selects the view with higher IoU, and Meaningful-gap Accuracy is additionally reported for pairs satisfying $|\Delta\text{IoU}| > 0.05$. Since the three measures have different numerical scales, they are compared at matched activation coverage.

As shown in Table VI, maximum response achieves the highest teacher-selection accuracy of 65.98%, compared with 49.26% for normalized entropy and 64.31% for APCE. Its Meaningful-gap Accuracy reaches 74.47%, also exceeding normalized entropy (52.30%) and APCE (72.70%). These results verify that maximum response provides an effective and lightweight relative reliability cue for CMD. Moreover, confidence-gap gating with $\delta = 0.05$ improves teacher-selection accuracy from 62.86% to 65.98% and Meaningful-gap Accuracy from 69.97% to 74.47%, demonstrating the benefit of filtering samples with an ambiguous teacher direction.

H. Ablation Study of VLVP

1) Overall Ablation Analysis: To investigate the respective contributions of visual and textual prompts within the VLVP module, we conduct a detailed ablation study, and the results are reported in Table III. First, removing the VLVP module entirely leads to a clear performance drop, which confirms the necessity of guiding the learning of view-specific feature representations. Second, using only visual prompts yields considerable improvements, suggesting that visual context plays an important role in fine-grained appearance alignment. Similarly, using only textual prompts also brings consistent gains, indicating that semantic viewpoint cues can effectively guide feature representation learning. Most importantly, the full model, which combines both visual and textual prompts, achieves the best performance across all metrics. This result demonstrates that detailed appearance information from visual features and high-level semantic context from textual prompts are complementary, thereby enabling more robust tracking under drastic view changes.

TABLE VII: Sensitivity analysis of the VLVP fusion weight α on the UGVT test set. "Learned" denotes the default setting, in which α is optimized jointly with the model.

α	UAV View		Ground View		Average	
	PR↑	SR↑	PR↑	SR↑	PR↑	SR↑
0.25	0.849	0.694	0.847	0.703	0.848	0.699
Learned	**0.886**	**0.717**	**0.853**	**0.704**	**0.869**	**0.711**
0.75	0.852	0.697	0.849	0.697	0.850	0.697
1.00	0.859	0.702	0.834	0.690	0.847	0.696

2) Comparison with Learnable Embedding: To further distinguish the complete VLVP from a simple view-identity encoding, we add a Learnable Embedding baseline that injects static UAV/ground view embeddings into PCVA through the same prompt interface, without CLIP semantics or target-specific visual context. As shown in Table III, Learnable Embedding obtains 0.869 and 0.706 in UAV PR and SR, respectively, compared with 0.886 and 0.717 for the complete VLVP. On the ground view, it obtains 0.846 and 0.699 in PR and SR, respectively, compared with 0.853 and 0.704 for the complete VLVP. Compared with Learnable Embedding, the complete VLVP achieves improvements of 1.7% and 1.1% in UAV PR and SR, respectively, and 0.7% and 0.5% in ground-view PR and SR, respectively. Together with the Only Visual Prompt and Only Text Prompt results, this shows that viewpoint-semantic priors and target-specific visual context provide complementary guidance beyond static view identity.

3) Sensitivity to Fusion Weight: In VLVP, the fusion weight α is a learnable scalar that is optimized jointly with the network, rather than a manually tuned hyperparameter. To explicitly examine its sensitivity, we additionally fix α to three representative values and evaluate the resulting models on the UGVT test set. As shown in Table VII, the learned setting achieves 0.869 and 0.711 in average PR and SR, respectively, outperforming all fixed alternatives. These results indicate that learning α allows VLVP to adaptively balance viewpoint semantics and visual context, yielding robust performance without manual tuning.

4) Effect of Viewpoint-Specific Language Guidance: To further verify the role of viewpoint-aware language prompts, we compare View-specific, Generic, and Swapped settings. In the Generic setting, both UAV and ground views use the prompt "an image of an object". In the Swapped setting, the original UAV and ground-view prompts are directly exchanged.

TABLE VIII: Ablation results of different language prompt settings on UGVT.

Prompt setting	UAV View		Ground View		Average	
	PR↑	SR↑	PR↑	SR↑	PR↑	SR↑
View-specific	**0.886**	**0.717**	**0.853**	0.704	**0.869**	**0.711**
Generic	0.863	0.700	0.814	0.677	0.839	0.688
Swapped	0.852	0.698	0.850	**0.706**	0.851	0.702

As shown in Table VIII, View-specific achieves the best average PR and SR. Compared with Generic, it achieves improvements of 3.0% and 2.3% in average PR and SR, respectively; compared with Swapped, the corresponding improvements are 1.8% and 0.9%. Generic removes the explicit viewpoint distinction from the language prompts, whereas Swapped disrupts the correspondence between the viewpoint semantics and their respective branches. These results show that viewpoint-specific language semantics and their correct branch correspondence contribute to tracking performance.

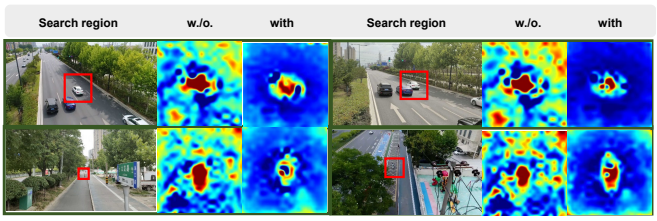


Fig. 8: Feature maps inside PCVA with and without prompts.

I. Efficiency Analysis

TABLE IX: Efficiency comparison under the same inference setting. Parameters are reported as trainable/total, while computation and latency are reported as initialization/online costs.

Method	Params Train./Total	FLOPs Init./Online	Latency Init./Online	Peak Mem.
Baseline	89.49M/89.49M	–/208.1G	–/7.05 ms	433.2 MB
Ours	98.67M/249.95M	17.5G/213.8G	4.10/7.98 ms	790.4 MB

To comprehensively evaluate the computational efficiency of VL-UniTrack, we further report the parameter count, FLOPs, model forward latency, and peak GPU memory in Table IX. The increase in peak memory from 433.2 MB to 790.4 MB mainly arises from the frozen CLIP ViT-B/32 model used to extract visual prompts. Although CLIP introduces 151.28M frozen parameters, the number of trainable parameters increases only from 89.49M to 98.67M, corresponding to an increase of 10.26%. Meanwhile, the online computation increases only from 208.1 GFLOPs to 213.8 GFLOPs, and the model forward latency increases from 7.05 ms to 7.98 ms. These results show that the additional memory is primarily associated with the frozen CLIP weights, while the increases in trainable model size, online computation, and inference latency remain limited. Online latency is measured using CUDA Events for model forward only after caching the initial CLIP prompts; it excludes image loading, crop generation, CPU preprocessing, host-to-device transfer, post-processing, and result writing, whereas the FPS reported in Table I includes the complete tracking pipeline.

J. Attribute-Level Evaluation

To further evaluate the robustness of VL-UniTrack under different challenging conditions, we conduct an attribute-level evaluation on the two challenge attributes provided with the released UGVT data, namely Full Occlusion and Out-of-View. All trackers are evaluated under the same one-pass evaluation protocol using PR and SR.

TABLE X: Performance comparison under different challenge attributes on UGVT. Results are reported as PR/SR; the best results are in **bold**.

Tracker	Full Occlusion	Out-of-View
UNTrack	0.663 / 0.500	0.883 / 0.665
XTrack	0.671 / 0.464	0.888 / 0.643
SSTrack	0.655 / 0.487	0.872 / 0.696
CRM*	0.583 / 0.395	0.597 / 0.451
TransMDOT*	0.635 / 0.472	0.881 / 0.714
MITracker*	0.713 / 0.510	0.825 / 0.636
VL-UniTrack	**0.768 / 0.552**	**0.957 / 0.798**

As shown in Table X, VL-UniTrack achieves the best PR and SR under both attributes. It obtains 0.768 and 0.552 in PR and SR, respectively, under Full Occlusion, achieving improvements over the strongest competing PR and SR of 5.5% and 4.2%, respectively. Under Out-of-View, it achieves 0.957 and 0.798 in PR and SR, respectively. When one view lacks reliable target observations, VL-UniTrack exploits complementary information from the other view to maintain stable tracking, demonstrating robustness to single-view failure.

K. Visualization Results

Tracking Result Visualization. To provide a more intuitive demonstration of the effectiveness of the proposed method, we visualize qualitative comparisons with the representative multi-view trackers CRM and TransMDOT, together with UNTrack and XTrack, as shown in Fig. 6. Across challenging scenarios involving occlusion, small targets, and background interference, our method maintains accurate target localization in both views and produces more stable tracking results than the compared methods. These qualitative comparisons further demonstrate the effectiveness of VL-UniTrack in exploiting cross-view complementarity for robust UAV-ground visual tracking.

Feature Visualization. To further demonstrate the effectiveness of our approach, we visualize the intermediate features during tracking and perform a t-SNE analysis, as shown in Fig. 7. It can be observed that, throughout the tracking process, even when the target appears extremely small, the responses of our method from both views remain consistently focused on the target. Moreover, the t-SNE results show that our method can clearly separate foreground targets from background regions. These analyses provide additional evidence that our method learns discriminative and view-consistent feature representations, thereby contributing to robust and stable tracking performance.

Visualization of the Role of Prompts in PCVA. To provide more direct evidence of the role of prompts in PCVA, we visualize the intermediate feature maps inside PCVA with and without prompts, as shown in Fig. 8. Without prompt guidance, the feature responses are more diffuse and are easily distracted by surrounding background structures. After introducing the prompts, the responses become more compact and concentrate more accurately on the target regions, with substantially reduced activation on irrelevant areas. This phenomenon is consistently observed across different scenes and in both UAV and ground views. These visualizations support our claim that the proposed prompts explicitly guide feature interaction in PCVA and help the network learn more discriminative view-aware representations.



Fig. 9: Visualization of Center Location Error (CLE) on challenging sequences, which displays the frame-by-frame error evolution of different trackers. The curves represent the Euclidean distance (in pixels) between the center of the predicted bounding box and the ground truth. A lower value indicates smaller location deviation and higher tracking accuracy.

L. Optimization Trajectory

To validate the stability of the proposed optimization strategy, we visualize the evolution of individual loss components together with the tracking performance (IoU) during training. As illustrated in Fig. 5, all loss terms, including the focal score-map loss, GIoU loss, L1 loss, and the proposed CMD loss, exhibit a consistent and steady downward trend. Notably, the convergence of the CMD loss indicates that mutual distillation effectively regularizes the unified feature space without disrupting the learning of the primary tracking tasks. Correspondingly, as the losses decrease, the IoU scores for both UAV and ground views consistently improve and eventually stabilize at a high level. These results confirm that the proposed multi-objective optimization formulation leads to robust tracking performance and ensures a stable training process.

M. Tracking Stability and Robustness Analysis

To further validate the robustness of VL-UniTrack, we select two challenging sequences and visualize the frame-level tracking deviation using the Center Location Error (CLE) metric. CLE is defined as the Euclidean distance between the predicted center and the ground-truth center, where a lower value indicates higher tracking accuracy.

As illustrated in Fig. 9, when the target undergoes severe occlusion or is distracted by similar objects in one view, as reflected by the spikes in the error curves, competing methods often exhibit large deviations or even tracking failure. In contrast, VL-UniTrack consistently maintains a low and

stable error curve throughout the sequences. This stability can be attributed to the cross-view complementarity enabled by PCVA: when one view is degraded, the network can effectively exploit reliable visual cues from the unobstructed counterpart to correct the prediction. These results demonstrate the effectiveness of cross-view collaboration in handling asymmetric degradation across views.

N. Failure Case Analysis

To provide a more comprehensive view of both the strengths and limitations of VL-UniTrack, we further analyze representative failure cases in Fig. 10. Under single-view degradation, such as an extremely small target, severe occlusion, or the target moving outside the field of view, the other view can retain a reliable observation, maintain stable localization, and assist the degraded view in recovering once its visual evidence returns. Temporary drift may occur in the more extreme case where both views simultaneously provide insufficient evidence, for example, when the target is extremely small in one view while being severely occluded or outside the field of view in the other. Once either view regains reliable target evidence, cross-view collaboration can correct the prediction and resume stable dual-view tracking. These cases demonstrate the stability and recovery capability of VL-UniTrack under single-view degradation, while the simultaneous loss of reliable observations in both views remains a challenging condition.

O. Discussion

The unified framework enables visual features from the UAV and ground views to interact sufficiently within a shared representation space, while viewpoint-specific cues also need to be fully represented. VLVP combines target appearance information from the initial templates with language-based viewpoint semantics to guide view-specific representations. The complete VLVP achieves higher average PR and SR than the Only Visual Prompt, Only Text Prompt and Learnable Embedding settings in Table III, as well as the Generic setting in Table VIII. These comparisons show that our combination of viewpoint semantics and target appearance information provides the most effective guidance among these alternatives, further improving tracking performance within the unified framework. In our tracking setting, the target is specified by its initial bounding boxes and visual templates, while language provides viewpoint-level semantic guidance.

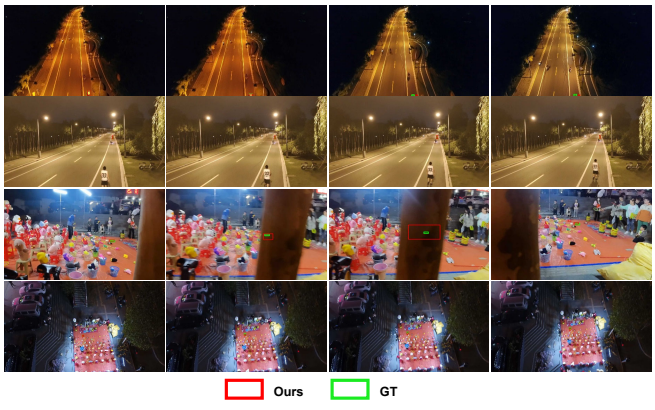


Fig. 10: Representative failure cases and recovery processes of VL-UniTrack under simultaneous cross-view degradation. Each pair of rows shows two synchronized views of the same sequence, and columns correspond to successive frames. Red and green boxes denote our predictions and the ground truth, respectively.

Specifying targets through language descriptions is another valuable capability. Representative studies, including Tracking by Natural Language Specification by Li *et al.* [38] and Siamese Natural Language Tracker [39], demonstrate the potential of language for expressing target semantics, supporting natural interaction, and complementing visual cues. Meanwhile, incorporating this capability into our framework requires consideration of potential inconsistencies between fixed object descriptions and dynamic visual observations: as the target's appearance or relative position changes, the original description may become inaccurate and interfere with localization. In future work, we will explore combining language-driven target specification with temporal visual evidence to improve the flexibility of target selection while maintaining stable UAV-ground collaborative tracking.

V. Conclusion

In this paper, we proposed VL-UniTrack, a unified framework for UAV-ground visual tracking. Specifically, we constructed a unified joint encoder to process dual-view inputs jointly, thereby reducing feature isolation between the two views and enabling sufficient cross-view interaction. To address the representation ambiguity caused by implicit appearance matching, we designed a visual-language viewpoint prompting module that fuses language descriptions with visual features to generate viewpoint-aware prompts. These prompts are further utilized by the prompt-guided cross-view adapter to explicitly guide cross-view feature interaction and the learning of view-specific representations. In addition, we introduced a confidence-modulated mutual distillation loss to stabilize training. Extensive experiments on the benchmark demonstrate that VL-UniTrack achieves state-of-the-art performance.

References

[1] Dengdi Sun, Leilei Cheng, Song Chen, Chenglong Li, Yun Xiao, and Bin Luo, "UAV-ground visual tracking: A unified dataset and collaborative learning approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3619–3632, 2024.

[2] Isuru Munasinghe, Asanka Perera, and Ravinesh C. Deo, "A comprehensive review of UAV-UGV collaboration: Advancements and challenges," *Journal of Sensor and Actuator Networks*, vol. 13, no. 6, Art. no. 81, 2024.

[3] Jianqiang Li, Genqiang Deng, Chengwen Luo, Qiuzhen Lin, Qiao Yan, and Zhong Ming, "A hybrid path planning method in unmanned air/ground vehicle (UAV/UGV) cooperative systems," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9585–9596, 2016.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 8748–8763.

[5] Sara Minaeian, Jian Liu, and Young-Jun Son, "Crowd detection and localization using a team of cooperative UAV/UGVs," in *Proceedings of the IIE Annual Conference and Expo*, 2015, pp. 595–604.

[6] Sara Minaeian, Jian Liu, and Young-Jun Son, "Vision-based target detection and localization via a team of cooperative UAV and UGVs," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 7, pp. 1005–1016, 2016.

[7] Yang Zhou, Jiuhong Xiao, Yue Zhou, and Giuseppe Loianno, "Multi-robot collaborative perception with graph neural networks," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2289–2296, 2022.

[8] Yuchao Wang, Zhirui Wang, Peirui Cheng, Pengju Tian, Ziyang Yuan, Jing Tian, Wensheng Wang, and Liangjin Zhao, "AVCPNet: An AAV-vehicle collaborative perception network for 3-D object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–16, 2025.

[9] Yunhao Hou, Bochao Zou, Min Zhang, Ran Chen, Shangdong Yang, Yanmei Zhang, Junbao Zhuo, Siheng Chen, Jiansheng Chen, and Huimin Ma, "AGC-Drive: A large-scale dataset for real-world aerial-ground collaboration in driving scenarios," 2025, arXiv:2506.16371.

[10] Pengfei Zhu, Jiayu Zheng, Dawei Du, Longyin Wen, Yiming Sun, and Qinghua Hu, "Multi-drone-based single object tracking with agent sharing network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4058–4070, 2021.

[11] Zhihao Liu, Yuanyuan Shang, Timing Li, Guanlin Chen, Yu Wang, Qinghua Hu, and Pengfei Zhu, "Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark," *IEEE Transactions on Multimedia*, vol. 25, pp. 1462–1476, 2023.

[12] Guanlin Chen, Pengfei Zhu, Bing Cao, Xing Wang, and Qinghua Hu, "Cross-drone transformer network for robust single object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4552–4563, 2023.

[13] Yuanliang Xue, Guodong Jin, Tao Shen, Lining Tan, Nian Wang, Jing Gao, and Lianfeng Wang, "Consistent representation mining for multi-drone single object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 10845–10859, 2024.

[14] Shengyu Hao, Peiyuan Liu, Yibing Zhan, Kaixun Jin, Zuozhu Liu, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang, "DIVOTrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes," *International Journal of Computer Vision*, vol. 132, no. 4, pp. 1075–1090, 2024.

[15] Huijie Fan, Tinghui Zhao, Qiang Wang, Baojie Fan, Yandong Tang, and LianQing Liu, "GMT: A robust global association model for multi-target multi-camera tracking," 2024, arXiv:2407.01007v1.

[16] Mengjie Xu, Yitao Zhu, Haotian Jiang, Jiaming Li, Zhenrong Shen, Sheng Wang, Haolin Huang, Xinyu Wang, Han Zhang, Qing Yang, and Qian Wang, "MITracker: Multi-view integration for visual object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 27176–27185.

[17] Hyeonseob Nam and Bohyung Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4293–4302.

[18] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte, "Learning discriminative model prediction for tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6182–6191.

[19] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proceedings of the European Conference on Computer Vision Workshops*, 2016, pp. 850–865.

[20] Bin Yan, Houwen Peng, Kan Wu, Dong Wang, Jianlong Fu, and Huchuan Lu, "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*, 2021, pp. 15180–15189.

[21] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen, "Joint feature learning and relation modeling for tracking: A one-stream framework," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 341–357.

[22] Liting Lin, Heng Fan, Zhipeng Zhang, Yong Xu, and Haibin Ling, "SwinTrack: A simple and strong baseline for transformer tracking," in *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 16743–16754.

[23] Xin Chen, Ben Kang, Jiawen Zhu, Dong Wang, Houwen Peng, and Huchuan Lu, "Unified sequence-to-sequence learning for single- and multi-modal visual object tracking," 2024, arXiv:2304.14394v3.

[24] Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei, "ARTrackV2: Prompting autoregressive tracker where to look and how to describe," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19048–19057.

[25] 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*, vol. 46, no. 12, pp. 9766–9779, 2024.

[26] 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*, 2025, vol. 39, no. 2, pp. 2239–2247.

[27] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu, "SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 7, pp. 12549–12556.

[28] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu, "Transformer tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8126–8135.

[29] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool, "Learning target candidate association to keep track of what not to track," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13444–13454.

[30] Zhihong Fu, Zehua Fu, Qingjie Liu, Wenrui Cai, and Yunhong Wang, "SparseTT: Visual tracking with sparse transformers," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2022, pp. 905–912.

[31] Yutao Cui, Tianhui Song, Gangshan Wu, and Limin Wang, "MixFormerV2: Efficient fully transformer tracking," in *Advances in Neural Information Processing Systems*, 2023, vol. 36, pp. 58736–58751.

[32] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool, "Transforming model prediction for tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8731–8740.

[33] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan, "AiATrack: Attention in attention for transformer visual tracking," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 146–164.

[34] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang, "Compact transformer tracker with correlative masked modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, no. 2, pp. 2321–2329.

[35] Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte, "Single-model and any-modality for video object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19156–19166.

[36] Yuedong Tan, Zongwei Wu, Yuqian Fu, Zhuyun Zhou, Guolei Sun, Eduard Zamfir, Chao Ma, Danda Paudel, Luc Van Gool, and Radu Timofte, "XTrack: Multimodal training boosts RGB-X video object trackers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 5734–5744.

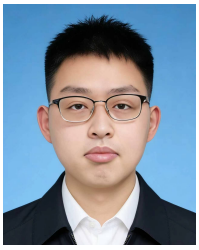
[37] Yaozong Zheng, Bineng Zhong, Qihua Liang, Ning Li, and Shuxiang Song, "Decoupled spatio-temporal consistency learning for self-supervised tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, no. 10, pp. 10635–10643.

[38] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees G. M. Snoek, and Arnold W. M. Smeulders, "Tracking by natural language specification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6495–6503.

[39] 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*, 2021, pp. 5851–5860.



Boyue Xu (Student Member, IEEE) received the M.S. degree from the School of Software Engineering, Nanjing University, Nanjing, China. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Nanjing University. His research interests include multi-modal representation learning, visual object tracking, and video object segmentation. He has published papers in journals and conferences, including Pattern Recognition, Image and Vision Computing, The Visual Computer, the IEEE International Conference on Multimedia and Expo, and the ACM International Conference on Multimedia Retrieval.



Ruichao Hou (Member, IEEE) received his Ph.D. degree from the Department of Computer Science and Technology, Nanjing University, in 2023. He is currently an Associate Professor with the School of Artificial Intelligence, China Pharmaceutical University. His research mainly focuses on multi-modal representation learning. He has published more than 40 papers in top-tier journals and conferences. He also serves as a reviewer for several prestigious journals, such as IEEE TPAMI, IEEE TNNLS, IEEE TMM, and IEEE TCSVT.



Tongwei Ren (Member, IEEE) received the B.S., M.E., and Ph.D. degrees from Nanjing University, Nanjing, China, in 2004, 2006, and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, China Pharmaceutical University, Nanjing, China. His research interest mainly includes multimedia computing and its real-world applications. He has published more than 100 papers in top-tier journals and conferences. He was a recipient of the best paper candidate awards of ICIMCS 2014, PCM 2015, ICMR 2020 and MMAsia 2020, and he was in the champion teams of ECCV 2018 PIC, MM 2019 VRU, MM 2020 DVU, MM 2022 DVU, and MM 2023 DVU challenges.



Gangshan Wu (Member, IEEE) received the B.Sc., M.S., and Ph.D. degrees from the Department of Computer Science and Technology, Nanjing University, Nanjing, China, in 1988, 1991, and 2000, respectively. He is currently a Professor with the School of Computer Science, Nanjing University. His current research interests include computer vision, multimedia content analysis, multimedia information retrieval, digital museum, and large-scale volumetric data processing.