MLLM-Track: An End-to-End Framework for Single Object Tracking

Hao Sun[1], Guosen Li[1], Mingzhe Zhang[1], Cun Ji[1], Xiangwei Zheng[1,*], Xinchun Cui[2]
{ 2024317216@stu.sdnu.edu.cn, 2024317199@stu.sdnu.edu.cn, zmz420@126.com
jicun@sdnu.edu.cn, xwzhengcn@163.com, cuixinchun@uhrs.edu.cn }
[1]School of Computer Science and Artificial Intelligence, Shandong Normal University, Jinan, China
[2]School of Foundational Education, University of Health and Rehabilitation Sciences, Qingdao, China
*Corresponding author

Abstract. Referring Single-Object Tracking (RSOT) requires locating objects via natural language. However, ambiguity in references and redundancy in visual features hinder performance. We propose MLLM-Track, an end-to-end framework. Its outer loop uses Reflective Prompt Optimization (RPO) to generate discriminative prompts via a Vision-Language Model (VLM), guided by unified alignment and localization scores. The inner loop features TGoT-K, which filters visual tokens via text-guided attention, and CM-GTR, a gated transformer using binary gating and sparse sampling to aggregate temporal cues efficiently. On the Elysium benchmark, MLLM-Track achieves 92.0% AUC, significantly outperforming baselines while maintaining fixed token budgets.

Keywords: RSOT, vision-language model, token selection, Qwen-VL, LLM.

1 Introduction

Referring Single-Object Tracking (RSOT) [1] requires continuously locating a single object in video based on natural language references. Compared to traditional RSOT transforms semantic conditions from bounding boxes and IDs to descriptive text, enabling stronger interactivity and interpretability. Concurrently, recent advancements in multimodal large language models (MLLM/VLM) have rapidly improved text-image alignment, localization, and video understanding, opening new avenues for ”text-to-vision” conditional tracking. For instance, models like CLIP [2] utilize alignment towers to construct robust text-image similarity spaces, while Qwen [3] series models perform high-resolution and long-video visual understanding. Fusion heads within these large models then handle inference and output.

Despite MLLM priors, RSOT faces three bottlenecks: 1) Textual ambiguity: Real-world references often omit attributes (e.g., ”the person on the left”), requiring refinement. 2) Visual redundancy: Full spatiotemporal attention propagates background noise. 3) Temporal noise propagation:

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365287

Lack of gating allows errors to accumulate over time. We propose MLLM-Track: an end-to-end large-model RSOT framework that structurally couples semantic and spatiotemporal paths through an "outer-loop text optimization, inner-loop sparsification and gating, lightweight fusion and prediction" architecture. The outer loop employs the Qwen-VL [3] model to generate candidate rewrites and diagnostics, selecting refined prompts via SigLIP [4] alignment and phrase-level localization scoring. The inner loop employs CLIP unified encoding, introducing TGoT-K for text-guided cross-modal attention and lightweight calibration within frames to filter high-scoring tokens. CM-GTR then applies binary gating to select tokens for cross-frame attention, with optional deformable attention aggregating key temporal clues at sparsely sampled points. Finally, it undergoes lightweight cross-attention fusion with text tokens before feeding into the Vicuna-7b-v1.5 fusion head for bounding box and mask prediction.

The main contributions of this paper are as follows:

- We propose a unified end-to-end large model RSOT framework (MLLM-Track), integrating "RPO outer loop, intra-frame Top-K compression, text-guided gated temporal and large model fusion" to achieve success AUC/Precision@20/Normalized-Precision AUC of 92.0/96.8/96.7(%), with Mean IoU of 92.1%.
- We develop TGoT-K (Text-Guided on-Token Top-K): a minimalist, interface-stable intra-frame compression module. It employs a single-head text-guided cross-modal attention average score, supplemented by lightweight MLP and EMA stabilization, to select the most relevant tokens for subsequent temporal modeling.
- We design CM-GTR (Cross-frame Masked Gated TRansformer): a binary gate selects tokens for cross-frame attention, while deformable sparse sampling aggregates key temporal information. During training, a differentiable approximation ensures end-to-end optimization, while inference provides deterministic gating.

2 Related Work

2.1 Evolution of RSOT Architectures

Early Referring Single-Object Tracking (RSOT) works [1, 5] demonstrated the feasibility of initializing trackers via natural language, evolving into standardized benchmarks like TNL2K and LaSOT [6, 7]. Architecturally, algorithms progressed from Siamese-based fusion [8] to unified "joint modeling" frameworks [9], which integrate grounding and tracking to handle appearance changes. On the pure vision side, Transformer-based trackers such as STARK [10] and OSTrack [11] have established robust spatiotemporal baselines. However, these "backbone with fusion" paradigms typically lack explicit, language-driven token selection mechanisms. Without effective gating, they are susceptible to propagating background noise, especially when visual cues are ambiguous or distractors are present.

2.2 MLLMs and Feedback Mechanisms

Recent advancements in MLLMs (e.g., CLIP [2], SigLIP [4], Qwen [3]) offer powerful priors for text-image alignment and video understanding. Despite their semantic strength, integrating them into RSOT under strict computational budgets remains challenging. The pioneering ChatTracker [12] introduced Reflective Prompt Optimization (RPO) to iteratively refine ambiguous descriptions using tracking feedback. While innovative, ChatTracker operates as a loosely coupled system, where language refinement and visual tracking are separate processes.

3 The Proposed Method

3.1 Overview

The overall approach of MLLM-Track is to couple reflective optimization on the language side with sparse spatiotemporal modeling on the visual side within a unified end-to-end framework. The system first employs a multimodal large model to diagnose key frames and initial reference text, generating multiple candidate descriptions to select a more discriminative refined prompt. This ensures the linguistic conditions align with video details while maintaining strong directionality. Subsequently, a unified image-text encoder maps each video frame to a visual token and the refined prompt to a text token. Within each frame, text-guided attention scoring [13], lightweight calibration, and temporal smoothing stably select the most relevant set of visual tokens, forming a fixed-size interface for the downstream temporal module. During cross-frame modeling, text-guided binary gating selects tokens for temporal attention. Sparse sampling aggregates information from key neighborhoods, suppressing temporal diffusion of irrelevant noise while maintaining robustness to occlusion recovery and large displacement scenarios. Finally, the gated temporal representations and text tokens are fused via a lightweight cross-attention, with a lightweight prediction head outputting bounding boxes and optional occlusion masks on the large model host. This design leverages recent systematic advances in unified image-text alignment, dynamic resolution, and long-video understanding: enhanced semantic understanding and localization in SigLIP; capabilities in dynamic resolution and long-video processing in Qwen2.5-VL; and the effectiveness of reflective prompt optimization in resolving real-world textual ambiguities. Compared to performing spatiotemporal attention directly on full visual tokens, this language-first pipeline maintains representational adequacy while significantly reducing redundant computations and noise propagation risks.

3.2 Feedback Optimization of Outer Loop

Using the key first frame as x_{t_0} and the initial reference text as $p^\star$. The objective is to obtain a reference $p^\star$ that aligns with the video and exhibits greater distinctiveness within a finite number of iterations. The method follows ChatTracker's reflective prompt optimization approach [14]: iteratively rewriting the reference based on tracking feedback to progressively resolve ambiguities and omissions through linguistic conditions while matching visual evidence.

By invoking the Qwen visual model combined with rules, a candidate set $\mathscr{P} = \{\tilde{p}^{(r)}\}_{r=1}^{R}$ is generated, where each $\tilde{p}^{(r)}$ explicitly captures attributes, locations, and relationships. f_v, f_w denote

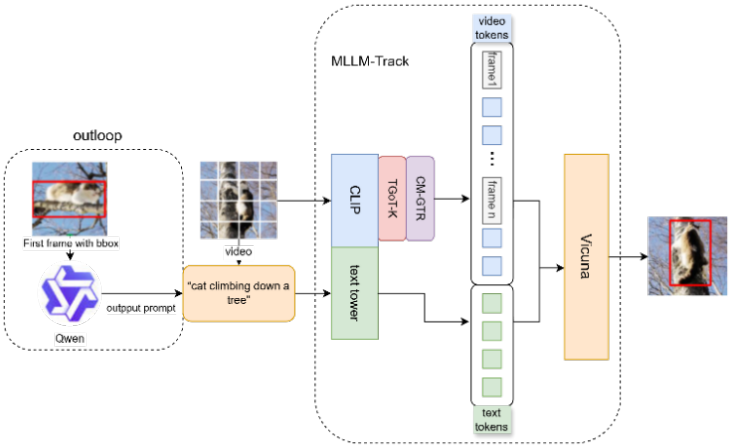

Fig. 1. The architecture of MLLM-Track: from prompt optimization to token filtering, gated temporal modeling, and fusion with a vision-language backbone.

SigLIP's visual and textual projections, $\mathrm{sim}(\cdot,\cdot)$ represents similarity in the projection space, $\Phi(\tilde{p})$ denotes the set of locatable phrases (noun phrases, attribute phrases, etc.) within candidates, and $\mathrm{score}_{\mathrm{ground}}(\phi \mid x_{t_0})$ is the localization score of phrase ϕ at the keyframe.

$$p^{\star} = \arg\max_{p \in P} \left(\mathrm{sim}(f_v(x_{t_0}), f_w(p)) + \lambda_g \max_{\phi \in \Phi(p)} \mathrm{score}_{\mathrm{ground}}(\phi \mid x_{t_0}) \right). \tag{1}$$

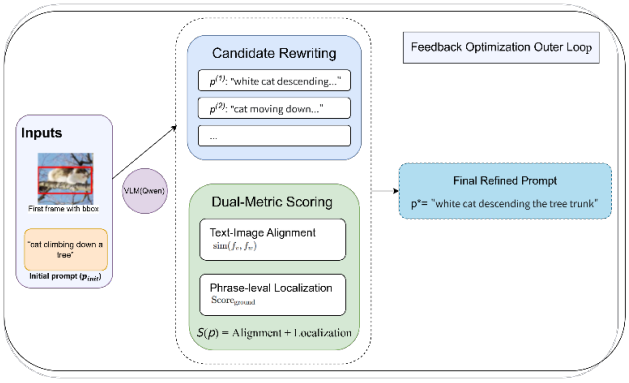

Fig. 2. Feedback Optimization Outer Loop, by invoking Qwen and performing computations, we obtain the optimal prompt.

Formula (1) combines "uniform alignment" and "phrase-level localizability" into a measurable criterion; $\lambda_g > 0$ serves as the weighting factor. The Qwen visual model delivers robust video understanding and localization capabilities, while CLIP [2] provides a stable image-text alignment space.

3.3 Text-Guided One-head Top-K (TGoT-K)

For any frame x_t, let the unified encoding yield the visual token matrix $V_t = \{v_{t,i}\}_{i=1}^{N} \in \mathbb{R}^{N\times d}$; for reference $p^\star$, let the text token matrix $W = \{w_j\}_{j=1}^{M} \in \mathbb{R}^{M\times d}$. Here, CLIP-ViT-Large-patch14 [2] is employed for visual tokens.

Subsequently, cross-modal attention scoring is performed. Define linear mappings $U_q \in \mathbb{R}^{d\times d_k}, U_k \in \mathbb{R}^{d\times d_k}$. To effectively identify visual regions corresponding to specific linguistic cues, we compute the attention distribution over visual tokens for each text token. The relevance score matrix is calculated as:

$$q_{t,i} = v_{t,i}U_q, \quad k_j = w_j U_k, \quad \alpha_{i,j} = \frac{\exp(\langle q_{i,i}, k_j\rangle/\sqrt{d_k})}{\sum_{n=1}^{N}\exp(\langle q_{i,n}, k_j\rangle/\sqrt{d_k})}, \tag{2}$$

where the Softmax normalization is applied along the visual token dimension N effectively generating a spatial heatmap for each word k_j. To determine the overall importance of each visual token, we aggregate the scores by taking the maximum value across the text dimension. This ensures that visual tokens highly aligned with distinct keywords (e.g., object names or attributes) are prioritized, regardless of the sentence length. The final selection score is defined as:

$$\bar{\alpha}_{t,i} = \max_j(\alpha_{i,j}), \quad \bar{\alpha}_{t,i} \in (0,1]. \tag{3}$$

The framework then performs lightweight calibration and fusion score calculation. Let $\bar{w} = \frac{1}{M}\sum_j w_j$, define a single-layer MLP $g(\cdot)$ with Sigmoid function $\sigma(\cdot)$ [9],

$$r_{t,i} = g([v_{t,i}; \bar{w}]) \in \mathbb{R}, \;\; s_{t,i} = \lambda_\alpha \bar{\alpha}_{t,i} + (1-\lambda_\alpha)\sigma(r_{t,i}), \lambda_\alpha \in [0,1]. \tag{4}$$

Subsequently, time smoothing and Top-K selection are implemented to ensure a constant interface [15]. To mitigate inter-frame jitter, an exponential moving average (EMA) is applied to the fusion score $s_{t,i}$ with a smoothing factor $\rho \in [0,1)$:

$$\hat{s}_{t,i} = \rho\hat{s}_{t-1,i} + (1-\rho)s_{t,i}. \tag{5}$$

Let Top-K return the set $S_t \subseteq \{1,\ldots,N\}$ of the token indices K with the highest scores, where $|S_t| = K$,

$$S_t = \mathrm{TopK}(\{\hat{s}_{t,i}\}_{i=1}^{N}, K), \;\; V_t^{\mathrm{sel}} = \{v_{t,i} | i \in S_t\} \in \mathbb{R}^{K\times d}. \tag{6}$$

Formula (6) establishes a stable "constant K interface" for downstream temporal modules, preventing dynamic length interference in batch processing and cross-frame attention.

3.4 Cross-frame Masked Gated Transformer (CM-GTR)

First, the framework incorporates a gated scoring mechanism and differentiable binarization to facilitate end-to-end feature selection. Let $V_t^{\mathrm{sel}} = \{v_{t,i}^{\mathrm{sel}}\}_{i=1}^{K}$ and w be as above. Define gated scoring as:

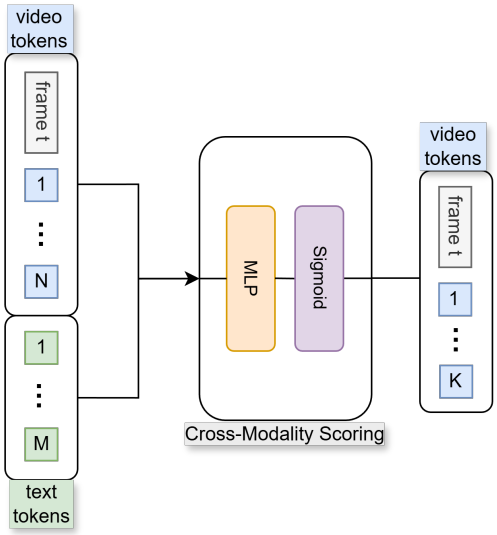

Fig. 3. Text-guided scoring of visual tokens per frame: select the top K most relevant tokens from the full set using cross-modality similarity and sigmoid-based calibration.

$$g_{t,i} = u_g^\top \mathrm{GELU}(W_v v_{t,i}^{\mathrm{sel}} + W_w \bar{w} + b_g) \in \mathbb{R}, \tilde{z}_{t,i} = \sigma(g_{t,i}) \in (0,1). \tag{7}$$

During training, Hard-Concrete continuous relaxation is employed to obtain differentiable gate switching with temperature $\beta = \frac{2}{3}$, stretching interval $(\gamma, \zeta) = (-0.1, 1.1)$, and $u \sim \mathrm{Uniform}(0,1)$,

$$s = \sigma\left(\frac{\log u - \log(1-u) + g_{t,i}}{\beta}\right), \tilde{z}_{t,i} = \mathrm{CLIP}(s(\zeta - \gamma) + \gamma, 0, 1). \tag{8}$$

The expected gate opening rate is used for sparse regularization (closed-form approximation). Let $m_{t,i}$ be the binary gate stochastic variable:

$$L_0 = \sum_{t,i} \Pr(m_{t,i} \neq 0) \approx \sum_{t,i} \sigma\left(g_{t,i} - \beta \log \frac{-\gamma}{\zeta}\right). \tag{9}$$

During inference, a fixed threshold $\tau = 0.5$ is used for "binarization": $m_{t,i} = 1[\tilde{z}_{t,i} \geq \tau]$, and the activation set is defined as $\mathscr{A}_t = \{i \mid m_{t,i} = 1\}$. The expectation and gradient propagation of Hard-Concrete can be found in the original paper.

Cross-frame attention is applied only to activated tokens. A two-layer multi-head attention is employed with hidden dimension $d_{\mathrm{hid}} = 512$ and heads $h = 8$. Layer l:

$$Q = V_{t,\mathscr{A}_t}^{\mathrm{sel}} W_q^{(l)}, K = V_{t-1,\mathscr{A}_{t-1}}^{\mathrm{sel}} W_k^{(l)}, V = V_{t-1,\mathscr{A}_{t-1}}^{\mathrm{sel}} W_v^{(l)}, Y_t^{(l)} = \mathrm{MSA}(Q,K,V), \tag{10}$$

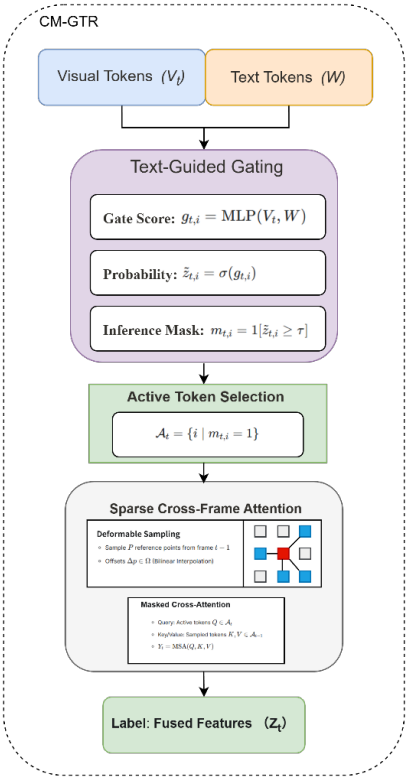

Fig. 4. Architecture of the proposed CM-GTR. The gated branch filters tokens based on linguistic-visual alignment, while the sparse cross-frame attention performs efficient temporal reasoning over the activated feature set.

where MSA denotes scaled dot-product multi-head attention (Transformer mechanism).

To take advantage of temporal dependencies, the framework performs sparse sampling of temporal aggregation [16]. Specifically, variable-form sparse sampling is employed in the temporal dimension to improve tracking robustness for small targets and scenarios involving large displacements. For each query i, sample from the adjacent frame offset set $\Omega = \{-2,-1,+1,+2\}$, where each offset takes $P = 4$ reference points. The weights and displacements are $A_{t,i,m}, \Delta p_{t,i,m}$ respectively, with the reference position being $p_{t,i}$. Define the bilinear interpolation operator $\Phi(\cdot)$, then

$$y_{t,i} = \sum_{m=1}^{P} A_{t,i,m} \Phi(V^{\mathrm{sel}}_{t+\Delta_m}, p_{t,i} + \Delta p_{t,i,m}), \Delta_m \in \Omega. \tag{11}$$

Formulas (10) and (11) confine attention to a small number of key neighborhoods, significantly reducing cross-frame computations and mitigating temporal noise propagation. The reference point aggregation concept for deformable attention originates from Deformable DETR [16].

3.5 Merging and Output

The gated temporal output undergoes a lightweight cross-attention with text tokens to explicitly aggregate linguistic conditions, yielding the fused representation Z_t. This is subsequently fed into the large model host's prediction head to output bounding boxes and optional masks:

$$Z_t = \text{CrossAttn}(Y_t, W), (b_t, m_t) = \text{Head}_{\text{VLM}}(Z_t). \tag{12}$$

This paper employs Vicuna-7b-v1.5 as the efficient host model. Its technical report and code demonstrate strong competitiveness in general visual understanding, multimodal reasoning, and interface localization tasks, while providing reproducible experiments and open-source weights.

3.6 Total Loss and Training Objective

The model is optimized end-to-end using a multi-task objective. The prediction head is supervised by L_{task}, which combines Generalized IoU loss [17] and L_1 loss for bounding box regression, along with Focal loss (L_{focal}) [18] and Dice loss (L_{dice}) [19] for mask segmentation.

To stabilize training, we incorporate two regularization terms: 1) Alignment Loss (L_{align}): Adopting InfoNCE [20], we align the text mean $\bar{w}$ with the frame-level visual embedding $\bar{v}_t$, calculated as the average of the selected token set V_t^{sel}. 2) Temporal Consistency Loss (L_{temp}): We maximize the cosine similarity between each selected token in frame t and its nearest neighbor in frame $t-1$, ensuring feature stability over time. The total objective is defined as:

$$L = L_{\text{task}} + \lambda_a L_{\text{align}} + \lambda_t L_{\text{temp}} + \lambda_0 L_0, \tag{13}$$

where λ terms are balancing hyperparameters.

4 Experiments

4.1 Evaluation Indicators

Following the standard One-Pass Evaluation (OPE) protocol [6, 7], we report three metrics: Success Score (AUC) calculated from the IoU overlaps, Precision (P) based on center location error (threshold at 20 pixels), and Normalized Precision (P_{norm}) to account for target scale variations.

4.2 Dataset and Baseline

The training and validation datasets utilize internally cleaned Elysium language-guided tracking data [21]. The training dataset utilized ElysiumTrack-1M, comprising approximately 1.27 million annotated video tracks. During training, ElysiumTrack-val500 served as the validation set for monitoring experimental performance. This validation set contained 500 independent video sequences, enabling real-time model evaluation after each training epoch. Performance metrics on the validation set guided critical hyperparameter tuning. We reproduce two practical reference categories: first, the "pure multimodal large model tracking" baseline (e.g., MiniGPT-v2-based approaches, serving as an upper bound for language alignment capability); second, the open-source implementation of an

existing RSOT pipeline ("Elysium baseline"), representing a strong, reproducible engineering baseline. All baselines utilize either the authors' publicly disclosed default settings or our validated optimal configurations to ensure fairness.

4.3 Experimental Settings

In addition to the training details outlined in §4, the following parameters are uniformly applied during the inference phase: input resolution 224×224, time window $T = 8$; K in TGoT-K is fixed at 32; CM-GTR threshold $\tau = 0.5$ with deformation sampling neighborhood $\{-2, -1, +1, +2\}$; all other hyperparameters remain consistent with training.

4.4 Results and Discussion

1) Performance Comparison: As shown in Table I, MLLM-Track achieves superior performance on the Elysium test partition, recording 92.0% Success (AUC), 96.8% Precision, and 96.7% Normalized Precision.

- **Comparison with Strong Baseline (Elysium):** Our method significantly outperforms the Elysium baseline (which scored 87.5% AUC). Specifically, MLLM-Track achieves absolute improvements of 4.50, 2.30, and 3.00 percentage points in AUC, Precision, and Normalized Precision, respectively. This represents a relative gain of 5.14% in AUC, verifying the effectiveness of our proposed token selection and gating mechanisms.
- **Comparison with MLLM Baseline (MiniGPT-v2):** The performance gap is even more substantial compared to the pure MLLM tracker MiniGPT-v2 (65.4% AUC). Our method leads by 26.60 percentage points in AUC (a relative gain of $\sim 40\%$), highlighting the necessity of our specialized inner-loop design over generic VLM inference.
- **Localization Accuracy:** In terms of overlap quality, MLLM-Track achieves a Mean IoU of 92.1%. This demonstrates that explicit linguistic conditions and controlled temporal propagation effectively enhance trajectory overlap and stability.

Table 1: RSOT Performance Comparison

Model	AUC (%)	P (%)	P_{Norm} (%)	IoU (%)
MiniGPT-v2	65.4	70.1	67.4	-
Elysium	87.5	94.5	93.7	-
MLLM-Track (ours)	**92.0**	**96.8**	**96.7**	**92.1**

2) Qualitative Component Analysis: We analyze the contribution of each module based on Table I. (i) Effectiveness of RPO: The significant gap between MiniGPT-v2 and MLLM-Track validates that RPO iteratively resolves textual ambiguity, concentrating alignment on the target. (ii) Necessity

of TGoT-K: This module is structurally essential for the "fixed token budget," effectively filtering irrelevant background noise via text-guided scoring before temporal modeling. (iii) Role of CM-GTR: The 4.5% Success gain over the Elysium baseline is primarily attributed to the gated transformer. Its binary gating selects only high-value tokens, while sparse aggregation ensures robust temporal propagation, preventing error accumulation during occlusion or fast motion.

3) Failure Case Analysis: Despite overall leading performance, we observe three failure scenarios. First, when initial descriptions contain contradictory or severely missing attributes, the outer loop may converge to suboptimal extremes during initial candidate generation, leading to intra-frame filtering errors. Second, following extreme camera cuts or prolonged full occlusions, the binary gate may excessively suppress newly generated tokens in the short term, causing recovery delays. Appropriately lowering the gate threshold can improve this but requires balancing false detections. Third, when ultra-small targets overlap complex high-frequency backgrounds, sparse sampling tends to aggregate around textures rather than the target itself. These observations align with public literature analyses: under the unified OPE protocol, Success and Precision metrics are highly sensitive to recovery speed, confirming that selective control over language and spatiotemporal aspects is crucial.

5 Conclusion

This paper proposes MLLM-Track for RSOT, achieving stable improvements in complex scenarios (occlusion, similar interference, fast motion) through closed-loop coupling of "language outer loop and vision inner loop" under constraints of fixed frame token budget and controllable temporal propagation. Specifically: (i) Reflective prompt optimization automatically resolves reference ambiguity under dual metrics of unified alignment and phrase-level localization, providing more discriminative semantic conditions for subsequent modeling; (ii) TGoT-K establishes a constant interface within frames via a minimalist mechanism, significantly reducing redundancy and enhancing temporal modeling stability; (iii) CM-GTR employs text-guided binary gating and reference-point sparse aggregation to effectively suppress irrelevant attention diffusion in the temporal dimension while preserving occlusion recovery and long-range dependencies under large displacements. Comprehensive experiments demonstrate that our method achieves significant advantages over strong baselines in Success/Precision/N-Precision/IoU metrics without increasing overall computational budget.

Limitations and future directions include three key points. First, the outer loop may still be driven by erroneous feedback into suboptimal rewrites during early iterations; adaptive iteration schedules and uncertainty constraints can be designed without significantly increasing latency. Second, the binary gate may temporarily suppress new tokens after extreme scene transitions or prolonged full occlusions; exploring prior-based reactivation strategies and gate threshold auto-scheduling is warranted. Third, sparse sampling occasionally exhibits texture bias in scenarios with ultra-small targets and high-frequency backgrounds, which can be mitigated by combining learnable geometric priors with multi-scale reference points. Future work will extend this closed-loop approach to multi-object language tracking and video open-vocabulary guidance scenarios, while integrating stronger unified encoding and localization capabilities and more robust discrete gate training strategies. This aims to unify efficiency and robustness within an end-to-end framework, with further cross-benchmark

reproducibility and energy-efficiency evaluations to be reported.

Acknowledgments

This work is supported by the Natural Science Foundation of Shandong Province China (NO. ZR2022LZH003, ZR2020LZH008),the Key R&D Program of Shandong Province, China (NO. 2021SFGC0104),the Undergraduate Teaching Reform Research Project of Shandong Province (BKJG2025211).

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepSeek in order to perform grammar and spelling checks. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] Li Z, Tao R, Gavves E, Snoek CGM, Smeulders AWM. Tracking by Natural Language Specification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 7350-8.

[2] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning(ICML). PMLR; 2021. p. 8748-63.

[3] Bai S, Chen K, Liu X, Wang J, Ge W, Song S, et al. Qwen2.5-VL Technical Report. arXiv preprint arXiv:250213923. 2025.

[4] Zhai X, Mustafa B, Kolesnikov A, Beyer L. Sigmoid Loss for Language Image Pre-Training. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023. p. 11941-52.

[5] Wang X, Shu X, Zhang Z, Jiang B, Wang Y, Tian Y, et al. Towards More Flexible and Accurate Object Tracking with Natural Language: Algorithms and Benchmark. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 13758-68.

[6] Fan H, Lin L, Yang F, Chu P, Deng G, Yu S, et al. LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 5369-78.

[7] Zhou L, Zhou Z, Mao K, He Z. Joint Visual Grounding and Tracking with Natural Language Specification. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 23151-60.

[8] Feng Q, Ablavsky V, Bai Q, Sclaroff S. Siamese Natural Language Tracker: Tracking by Natural Language Descriptions with Siamese Trackers. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 5847-56.

[9] Wu D, Han W, Wang T, Dong X, Zhang X, Shen J. Referring Multi-Object Tracking. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 14633-42.

[10] Lv H, Zhou C, Cui Z, Xu C, Li Y, Yang J. Localizing Anomalies From Weakly-Labeled Videos. IEEE Transactions on Image Processing. 2021;30:4505-15.

[11] Xu Z, Xu C, Cui Z, Zheng X, Yang J. CVNet: Contour Vibration Network for Building Extraction. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 1373-81.

[12] Sun Y, Yu F, Chen S, Zhang Y, Huang J, Li Y, et al. ChatTracker: Enhancing Visual Tracking Performance via Chatting with Multimodal Large Language Model. Advances in Neural Information Processing Systems (NeurIPS). 2024;37:39303-24.

[13] Liang Y, Ge C, Tong Z, Song Y, Wang J, Xie P. Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. In: International Conference on Learning Representations (ICLR); 2022. p. 1-10.

[14] Zhang L, Zheng X, Chen X, Cui L. Three-Dimensional View Relationship-Based Context-Aware Emotion Recognition. IEEE Transactions on Neural Networks and Learning Systems. 2025;36(7):13567-78.

[15] Rao Y, Zhao W, Liu B, Lu J, Zhou J, Hsieh CJ. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In: Advances in Neural Information Processing Systems. vol. 34; 2021. p. 13937-49.

[16] Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In: International Conference on Learning Representations(ICLR); 2021. p. 1-10.

[17] Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. p. 658-66.

[18] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. vol. 42; 2020. p. 318-27.

[19] Zheng X, Zhang L, Xu C, Chen X, Cui Z. An attribution graph-based interpretable method for CNNs. Neural Networks. 2024;179:106597.

[20] van den Oord A, Li Y, Vinyals O. Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:180703748. 2018.

[21] Wang H, Wang Y, Ye Y, Nie Y, Huang C. Elysium: Exploring Object-level Perception in Videos via MLLM. Proceedings of the European Conference on Computer Vision (ECCV). 2024:166-85.