This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

MVLM: Template-Free Tracking via Vision–Language Margin Confidence and Memory-Gated Tracking

Dae-Hyeon Park Mina Baek Jeong-Hun Ha Chan-Seop Park
Jamshidjon Ganiev Seung-Hwan Bae*
Inha University, Dept. of Electrical and Computer Engineering, Vision and Learning Lab.
{saintPalite2221,dkdend123,jhha,pcs1728,jamshid24}@inha.edu, shbae@inha.ac.kr

https://mvlm-tft.github.io

Abstract

We introduce a new template-free tracking paradigm based solely on natural language, capable of tracking an arbitrary object and seamlessly switching to a new target without box initialization. Our key idea is to localize an object via vision-language (VL) correlation. However, using the correlation alone is brittle under large search regions due to spatial uncertainty and ambiguous VL saliency. To resolve these, we propose MVLM, a memory-based vision-language margin confidence that integrates vision–language correlation, encoder prediction, and temporal memory. MVLM dynamically gates the search region—switching between compact ROI (Region of Interest) search and global re-localization—to reduce spatial uncertainty. Theoretically, we derive bounds that connect the MVLM score to tracking probability, characterizing mis-localization within ROI and ROI-exclusion probabilities. Through extensive evaluation, we validate our theorems and achieve state-of-the-art performance on several benchmarks (TNL2K, LaSOT, OTB99, and MGIT) using only language guidance.

1. Introduction

Template-free (TF) tracking with natural language eliminates the need for visual templates or box initialization, allowing a tracker to follow arbitrary objects and seamlessly switch to a new target by updating the text description. This is beneficial for open-world and interactive scenarios, where the user may not have an initial bounding box and the target can change on the fly. Although the earlier works [23, 26, 30, 38] state language-based tracking, they exploit object templates as an input to a tracker. In contrast, our approach eliminates visual object templates completely, relying solely on vision-language correlation as the semantic anchor.

A natural approach is to localize the object by measuring the correlation between vision features (from a search image) and a language query. However, correlation-alone tracking is unstable under broad search regions: spatial uncertainty grows with the area to scan, and vision–language (VL) saliency becomes ambiguous in the presence of distractors, occlusions, or appearance shifts. As a result, trackers that rely only on instantaneous VL correlation tend to mislocalize or lose the target when the search region is large. Therefore, recent transformer-based object trackers [3–8, 13, 15, 31–34, 36] employ local ROI search for restricting the search region around the tracked object predicted at a past frame. From these observations, the main problem of TF tracking can be transformed into learning distinctive saliency from vision-language correlation and determining object search regions around an object or global image region according to tracking and language guidance.

To resolve these, we propose memory-based vision-language confidence (MVLM) by incorporating three complementary features: (i) a VL correlation margin that separates the best-matching hypothesis from its nearest competitors, (ii) the encoder's output statistics that capture local visual evidence, and (iii) a lightweight temporal memory that stabilizes decisions across frames. We then exploit MVLM to gate the search strategy: when MVLM confidence is high, the tracker restricts itself to a compact ROI around the current estimate; when confidence drops, it triggers global re-localization over the whole image. This memory-gated tracking reduces spatial uncertainty without sacrificing recoverability after failures or during target switching.

Beyond the algorithm design, we provide a theoretical analysis that links MVLM to tracking success. In particular, we derive bounds that connect the MVLM score to the

*Corresponding author (shbae@inha.ac.kr).

probability of correct object localization, decomposing failure cases into two interpretable terms: (i) *mis-localization within the ROI* (incorrect ranking among candidates inside the restricted region) and (ii) *ROI exclusion* (omitting the ground-truth box due to the tight search region). These results formalize our intuition that enlarging the correlation margin and selecting an appropriate region size together suppress failure rates exponentially, and they justify MVLM confidence as a principled mechanism for switching between local and global search.

Compared to template-based trackers [11, 17, 18, 23, 26, 27, 30, 38], our approach removes the necessity of an initial visual exemplar and avoids error accumulation from stale templates. Compared to language-guided retrieval or visual grounding methods [16, 21, 35, 38], our method explicitly tackles the temporal nature of tracking with a memory that regularizes frame-wise decisions and a gating rule that adapts the search scope online. Practically, MVLM is computation-light, adds negligible overhead to a transformer-based backbone [37], and can be integrated with standard training objectives via a margin-aware loss.

To the best of our knowledge, this is the first TF tracking work based on a transformer architecture. Extensive experiments on LaSOT [9], TNL2K [29], OTB99 [20], and MGIT [14] validate our theory and achieve state-of-the-art (SOTA) performance using only language guidance. We further demonstrate the unique capability of our approach through target switching experiments, where the tracker successfully adapts to new language descriptions mid-sequence.

The main contributions of this work are:

- Introducing a template-free, language-only tracking paradigm that follows arbitrary objects and *seamlessly switches* to new targets without box initialization, controlled purely by natural language.
- Proposing an MVLM confidence that fuses correlation margins, encoder predictions, and temporal memory, and we leverage it to realize memory-gated tracking–dynamically switching between compact ROI search and global re-localization.
- Developing a theoretical framework relating MVLM thresholds to tracking success, yielding bounds that characterize mis-localization within the ROI and ROI-exclusion.
- Achieving state-of-the-art results on TNL2K, LaSOT, OTB99, and MGIT under language-only guidance.

In summary, MVLM turns raw VL correlation into a reliable, temporally consistent confidence measure and uses it to adapt the search region online. This closes the gap between language-guided localization and robust tracking, enabling template-free operation and smooth target switching in challenging, open-world scenarios.

2. Related works

Conventional single-object tracking follows a template-based paradigm [8, 36], where a cropped exemplar from the initial frame is used as a visual template for localization. Siamese networks [8] adopt this design by extracting template features from the initial frame and matching them with search region features. It relies on the quality and persistence of the template. Once the target undergoes appearance variations or becomes occluded, the tracker often fails. Above all, cropping a template often requires manual annotation at the moment of appearance, limiting practical human-interactive tracking. To avoid limitations, recent visual-language tracking frameworks integrate natural language descriptions as auxiliary cues, yet these approaches still depend on visual templates and therefore remain only partially human-interaction friendly. For example, GTI [35] and QueryNLT [26] use natural-language only for initialization, after which tracking proceeds in a standard template-based manner. JointNLT [38] and UVLTrack [23] unify grounding and tracking in a transformer but still depend on the initial grounded box as a visual reference. STMTrack [12] replaces the template with a label map, which still explicitly provides the target region. Some trackers [23, 38] use visual grounding to localize the target from language, but grounding performs short-term correspondence between language and image. Furthermore, grounding-based approaches [16, 21] inherently rely on region proposals by object detectors [1, 2] In summary, to the best of our knowledge, the study focusing on template-free tracking is still insufficient. We provide both theoretical formulation and a tracking framework under template-free setting.

3. Proposed methods

3.1. Basic formulation for VL correlation

Template-free tracking localizes the target without any visual exemplar or human box annotation. To resolve this, we localize the object using statistical correlations between visual search tokens and the given language embeddings. We formalize a TF tracking problem to identify the maximal correlation in a search image. Let $F \in \mathbb{R}^{H\times W\times C}$ be a feature map of a search image. We define the spatial index set $\Omega = \{(i,j) : 1 \le i \le H,\ 1 \le j \le W\}$, where $H = W = \sqrt{N_I}$. For $x \in \Omega$, the vision token embedding is $\mathbf{v_x} \equiv V_{(i,j)} \in \mathbb{R}^C$. Here, the subscript $\mathbf{x} = (i,j)$ means a vision token location within an image grid. We define the spatial center of token (i,j) in the input-image coordinate system by $\mathbf{c}(i,j) = \big(c_x(i,j),\, c_y(i,j)\big) = \big((j-\frac{1}{2})\,s,\ (i-\frac{1}{2})\,s\big)$ with pixel stride s of the feature map. For a candidate box $b = [x_1, y_1, x_2, y_2]$, its token index set is $R(b) = \big\{(i,j) \in \Omega :\ x_1 \le c_x(i,j) < x_2,\ \ y_1 \le c_y(i,j) < y_2\big\}$ (*i.e.* tokens fall inside b). We denote the ground-truth box by b^* and its token set by $R^* = R(b^*)$.

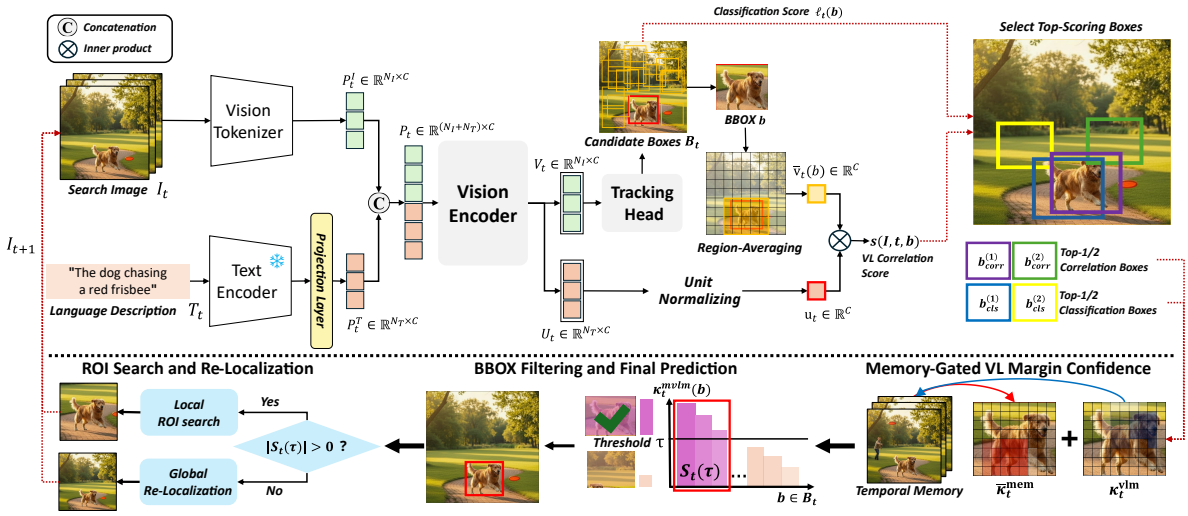


Figure 1. The illustration of our template-free tracking framework.

Given a unit-normalized language embedding $\mathbf{u}_t \in \mathbb{R}^C$ and vision tokens $\mathbf{v_x}$, the region-averaged alignment score is

$$s(I,t,b) = \frac{1}{|R(b)|} \sum_{\mathbf{x} \in R(b)} \langle \mathbf{v_x}, \mathbf{u}_t \rangle := \langle \bar{\mathbf{v}}_t(b), \mathbf{u}_t \rangle \quad (1)$$

To analyze the reliability of the alignment score under noise, we introduce a sub-Gaussian noise model that decomposes each observed $\mathbf{v_x}$ as: $\mathbf{v_x} = \boldsymbol{\mu}_\mathbf{x} + \boldsymbol{\varepsilon}_\mathbf{x}$ where $\boldsymbol{\varepsilon}_\mathbf{x}$ is zero-mean and $\langle \boldsymbol{\varepsilon}_\mathbf{x}, \mathbf{u}_t \rangle$ is sub-Gaussian with proxy variance σ^2. We then define the region mean vector and correlation for a box b with region $R(b)$ as: $\bar{\boldsymbol{\mu}}(b) = \frac{1}{|R(b)|} \sum_{\mathbf{x} \in R(b)} \boldsymbol{\mu}_\mathbf{x}$, and its alignment score $\rho(b) = \langle \bar{\boldsymbol{\mu}}(b), \mathbf{u}_t \rangle$ with the language embeddings.

3.2. Template-free tracking architecture

This design choice enables direct interaction of visual and language tokens, promoting stable and simplified training.

Overview. Our architecture comprises a frozen text encoder, a transformer-based vision encoder, and a three-branch prediction head for classification, size regression, and offset regression. For text encoding, we adopt the pre-trained CLIP [25] text encoder and freeze its parameters during training. Given a natural-language description T_t at frame t, we tokenize it and pass through the text encoder to generate a sequence of text tokens $P_t^T \in \mathbb{R}^{N_T \times C}$. For visual tokenization, the input image I_t is passed through a Conv2d layer and layer normalization to produce $P_t^I \in \mathbb{R}^{N_I \times C}$. After that, we produce a unified token sequence $P_t = [P_t^I; P_t^T] \in \mathbb{R}^{(N_I+N_T)\times C}$ by concatenating P_t^I and P_t^T. Then P_t is processed by a stack of transformer encoder layers [28], where alternating self-attention and feed-forward blocks enable cross-modal interaction and iteratively refine token representations. After joint encoding, we obtain the refined embeddings $P_t^* = [V_t; U_t]$, where $V_t = \{\mathbf{v_x}\}_{\mathbf{x} \in \Omega} \in \mathbb{R}^{N_I \times C}$ denotes the set of refined visual tokens in Ω and $U_t \in \mathbb{R}^{N_T \times C}$ denotes the set of refined language tokens. We define $\mathbf{u}_t$ as the mean-pooled language embedding: $\mathbf{u}_t = \frac{1}{N_T} \sum_{\mathbf{u} \in U_t} \mathbf{u}$, which represents the unified semantic description of the target. The V_t is fed to the prediction heads to generate the tracking outputs. During inference, the classification, offset, and size heads are aligned on the same spatial grid Ω. Therefore, each grid location (i,j) outputs a matched set of predictions as: classification score, an offset, and a size. Each grid position maps to an image-space center via a stride of $\sqrt{N_I}$, and combining this center with the predicted offset and size yields a candidate bounding box $b_{i,j}$. We define the candidate set as $\mathcal{B}_t = [b_{i,j} | (i,j) \in \Omega]$. The final box is selected from $\mathcal{B}_t$ according to Section 3.3.

Correlation margin loss. To train our template-free tracker, we design a Correlation Margin (CM) loss that directly maximizes the correlation margin $\Delta\rho$ between the ground-truth and background regions, motivated by the exponential decay structure of Theorem 1. Let $R^* = R(b^*)$ denote vision tokens indices inside the ground-truth box b^* and R^{neg} the background tokens outside the ground-truth box. We compute ρ_{pos} as the mean correlation over R^*:

$$\rho_{pos} = \frac{1}{|R^*|} \sum_{\mathbf{x} \in R^*} \langle \mathbf{v_x}, \mathbf{u}_t \rangle, \quad (2)$$

and define the margin to each background token as $\Delta\rho_\mathbf{x} = \rho_{pos} - \langle \mathbf{v}_x, \mathbf{u}_t \rangle$. Then, our loss minimizes:

$$\mathcal{L}_{CM} = \frac{1}{K} \sum_{\mathbf{x} \in \mathrm{TopK}(R^{neg})} \exp\left(-\tau \Delta\rho_\mathbf{x} |\Delta\rho_\mathbf{x}|\right), \quad (3)$$

where $\tau > 0$ is temperature and Top-K selects the K hardest negatives (smallest margins). When $\Delta\rho_\mathbf{x} > 0$, the loss

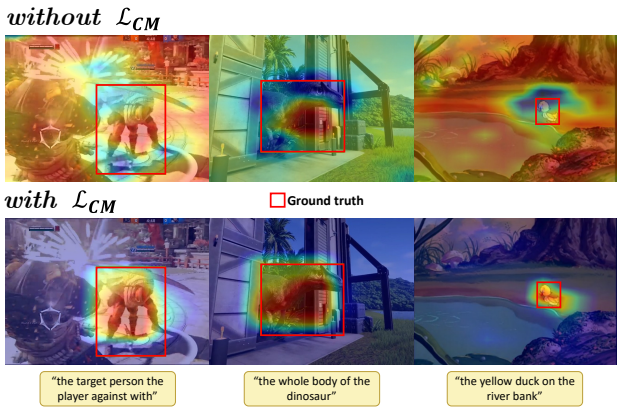


Figure 2. Attention visualization with/without the correlation margin loss $\mathcal{L}_{CM}$.

is small (good separation); when $\Delta\rho_{\mathbf{x}} < 0$, the exponential penalty grows rapidly. By minimizing $\mathcal{L}_{CM}$, we directly reduce the mis-localization bound in Theorem 1. As shown in Figure 2, $\mathcal{L}_{CM}$ concentrates correlation on the target region, sharpening vision-language alignment.

Overall training loss. We adopt heatmap-based prediction heads and a composite training loss following [7], where three task heads and their corresponding task losses are used. Our tracker is trained by the total loss $\mathcal{L}_{total} = \mathcal{L}_{track} + \mathcal{L}_{CM}$, where $\mathcal{L}_{track}$ is implemented by [36]. More details are mentioned in Supplementary Materials.

3.3. Memory VLM confidence for re-localization

Since tracking with VL correlation alone is unstable under distractors and appearance shifts, we propose MVLM confidence to dynamically adapt the search region and stabilize tracking decisions over time.

Vision-language margin confidence. Let $\mathcal{B}_t$ be the set of candidate boxes at frame t. Define the top-1 correlation box as $b^{(1)}_{\text{corr}} := \arg\max_{b\in\mathcal{B}_t} s(I,t,b)$. Its *outside* candidate pool is $\mathcal{B}^{\text{out}}_t(b) := \{b' \in \mathcal{B}_t : \text{IoU}(b',b) \le \psi_{out}\}$, where an exclusion threshold $\psi_{out} \in [0,1)$. The correlation-based top-2 is then selected *outside* the top-1 region: $b^{(2)}_{\text{corr}} := \arg\max_{b\in\mathcal{B}^{\psi_{out}}_t(b^{(1)}_{\text{corr}})} s(I,t,b)$. Choosing a small ψ_{out} (e.g., 0.1) means a strong distractor could exist outside the object region. Similarly, let $\ell_t(b) \in \mathbb{R}$ be the classification score for each candidate box b at frame t, which is predicted by the classification head. Let $b^{(1)}_{\text{cls}} = \arg\max_{b\in\mathcal{B}_t} \ell_t(b)$ and $b^{(2)}_{\text{cls}} = \arg\max_{b\in\mathcal{B}^{\text{out}}_t(b^{(1)}_{\text{cls}})} \ell_t(b)$ be top-1/2 by $\ell_t(b)$. Then we introduce standardized margins:

$$\tilde{\Delta}^{\text{corr}}_t := \frac{s\big(I,t,b^{(1)}_{\text{corr}}\big) - s\big(I,t,b^{(2)}_{\text{corr}}\big)}{\hat{\sigma}_{\text{corr}}} \tag{4}$$

$$\tilde{\Delta}^{\text{cls}}_t := \frac{\ell_t\big(b^{(1)}_{\text{cls}}\big) - \ell_t\big(b^{(2)}_{\text{cls}}\big)}{\hat{\sigma}_{\text{cls}}}, \tag{5}$$

where $\hat{\sigma}_{\text{corr}}, \hat{\sigma}_{\text{cls}} > 0$ are robust spread estimates (e.g., per-frame MAD or running EWMA std). At frame t, we aggregate the correlation-based $\tilde{\Delta}^{\text{corr}}_t \in [-1,1]$ and the classification-based $\tilde{\Delta}^{\text{cls}}_t \in [-1,1]$ margins into a vision-language margin (VLM) confidence:

$$\kappa^{\text{vlm}}_t := \alpha_{\text{corr}}\,\tilde{\Delta}^{\text{corr}}_t + \alpha_{\text{cls}}\,\tilde{\Delta}^{\text{cls}}_t, \tag{6}$$

where $\alpha_{\text{corr}}, \alpha_{\text{cls}} \ge 0$ and $\alpha_{\text{corr}} + \alpha_{\text{cls}} = 1$ holds for the nonnegative weights α_{corr} and α_{cls}. Under the normalization of $\tilde{\Delta}^{\text{corr}}_t, \tilde{\Delta}^{\text{cls}}_t \in [-1,1]$, κ^{vlm}_t is bounded in $[-1,1]$. It can be interpreted as a measure of how much a distinct correlation and classification score exists at a single frame.

Memory-gated VLM confidence. To stabilize decisions over time, we use an exponentially weighted moving average (EWMA) with the previous VLM confidences up to frame t:

$$\bar{\kappa}^{\text{mem}}_t := \sum_{\ell=0}^{t}(1-\lambda)\lambda^{\ell}\,\kappa^{\text{vlm}}_{t-\ell}, \qquad \lambda \in [0,1). \tag{7}$$

Here λ is the forgetting factor: larger λ increases temporal smoothing (lower variance) at the cost of longer lag. The weights sum to one, so $\bar{\kappa}^{\text{mem}}_t$ is a convex average of previous κ^{vlm}. In practice we use the equivalent one-step recursion

$$\bar{\kappa}^{\text{mem}}_t = (1-\lambda)\,\kappa^{\text{vlm}}_t + \lambda\,\bar{\kappa}^{\text{mem}}_{t-1}, \bar{\kappa}^{\text{mem}}_0 := \kappa^{\text{vlm}}_1. \tag{8}$$

Finally we blend the instantaneous and memory confidences:

$$\kappa^{\text{mvlm}}_t = (1-\omega)\,\kappa^{\text{vlm}}_t + \omega\,\bar{\kappa}^{\text{mem}}_t,\ \omega \in [0,1]. \tag{9}$$

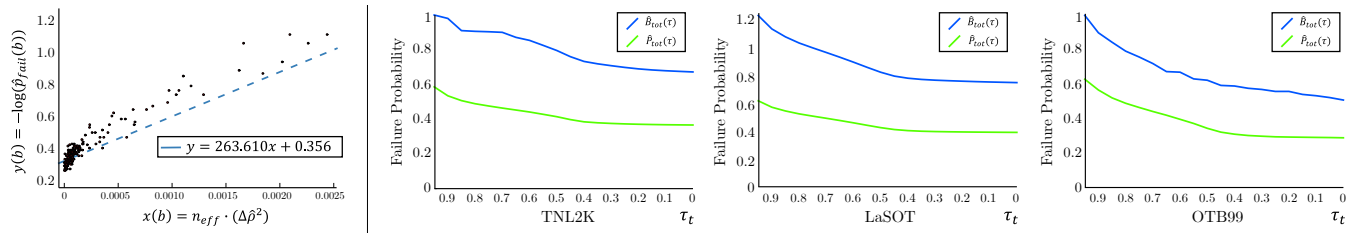
Setting $\omega = 0$ yields purely instantaneous decisions, while $\omega = 1$ relies entirely on the temporally smoothed memory. Under the boundedness assumptions above, κ^{mvlm}_t is bounded in $[-1,1]$ as well.

MVLM confidence-driven re-localization. Given a composite confidence κ^{mvlm}_t, we define the ROI selection as the subset of $\mathcal{B}_t$ whose per-box MVLM confidence exceeds τ:

$$\mathcal{S}_t(\tau) := \{\, b \in \mathcal{B}_t : \kappa^{\text{mvlm}}_t(b) \ge \tau \,\} \subseteq \mathcal{B}_t. \tag{10}$$

Here $\mathcal{S}_t(\cdot)$ can be considered as selected candidate boxes using κ^{mvlm}_t. For each candidate box $b \in \mathcal{B}_t$, we treat b as $b^{(1)}_{\text{corr}}$ and $b^{(1)}_{\text{cls}}$, and compute a per-box confidence $\kappa^{\text{mvlm}}_t(b)$. For comparisons between candidates within the ROI, we use the classification head output scores. The final bounding box $\hat{b}_t$ is determined as follows:

- If $|\mathcal{S}_t| > 0$: Select $\hat{b}_t = \arg\max_{b\in\mathcal{S}_t} \ell_t(b)$ and crop the search region around the predicted target for frame $t+1$ to enable local search.
- If $|\mathcal{S}_t| = 0$: Select $\hat{b}_t = \arg\max_{b\in\mathcal{B}_t} \ell_t(b)$ and use the full image as the search region for frame $t+1$ to enable global re-localization.



(a) Mis-localization probability

(b) Tracking failure probability

Figure 3. (a) Validation of Theorem 1: Collapse plot showing exponential decay of mis-localization probability. (b) Empirical validation of Theorem 2. The measured total failure rate $\hat{p}_{tot}(\tau)$ (green solid line) is plotted against the theoretical upper bound $\hat{B}_{tot}(\tau)$ (blue solid line) across a sweep of τ. In all dataset cases, $\hat{p}_{tot}(\tau)$ remains below $\hat{B}_{tot}(\tau)$.

4. Theoretical analysis

4.1. Correlation margin and mis-localization bound

An important question is whether the target is localized using only VL correlation. Intuitively, successful localization requires that the target region exhibits sufficiently stronger correlation with the language query than background regions. We formalize this intuition and derive probabilistic bounds characterizing when correlation-based tracking succeeds.

Recall from Sec. 3.1 that for a candidate box b, the *correlation margin* $\Delta\rho(b) = \rho(b^*) - \rho(b)$ between the GT box b^* and distractor b. We assume there exists a positive margin $\gamma > 0$ such that $\Delta\rho(b) \geq \gamma$ holds for all $b \neq b^*$. This ensures that the GT region maintains strictly higher alignment with the language description than any competing candidate. In practice, semantic precision in the language description and visual distinctiveness in the target jointly determine γ.

Theorem 1 (Mis-localization bound under sub-Gaussian noise)**.** *Under the assumptions on sub-Gaussian noise and correlation margin, the mis-localization probability is bounded as:*

$$\mathbb{P}\Big[\max_{b\neq b^*} s(I,t,b) \;\geq\; s(I,t,b^*)\Big] \leq \sum_{b\neq b^*} \exp\left(-\frac{\big(\rho(b^*)-\rho(b)\big)^2}{2\sigma^2\left(\frac{1}{|R(b^*)|}+\frac{1}{|R(b)|}\right)}\right). \tag{11}$$

Theorem 1 reveals that mis-localization probability decays *exponentially* with the correlation margin $\Delta\rho(b)$, justifying our Correlation Margin (CM) loss design (Section 3.2) that directly maximizes $\Delta\rho(b)$ during training. It derives the result that enough correlation margin is equivalent to the reliability of vision–language (VL) correlation.

4.2. Overall failure bound under MVLM confidence

Although Theorem 1 shows that sufficient correlation margin ensures reliable localization, distractors or appearance shifts can reduce this margin, leading to two failure modes: (i) mis-localization within ROI and (ii) ROI-exclusion. We formalize this via the following bound.

Theorem 2 (MVLM confidence re-localization bound)**.** *Under the sub-Gaussian noise model with a common token-level proxy variance σ^2 inside the ROI, assume:*

(a) Equal region size: $|R(b)| = n, \quad \forall b \in \mathcal{S}_t(\tau)$

(b) Bounded ROI exclusion: $\mathbb{P}\big[\, b^* \notin \mathcal{S}_t(\tau) \mid \tau \,\big] \;\leq\; \eta(\tau)$

(c) Minimum margin: $\rho(b^*) - \rho(b) \;\geq\; \gamma(\tau), \quad \forall b \in \mathcal{S}_t(\tau) \setminus \{b^*\}$

(d) Finite candidates: $M(\tau) := |\mathcal{S}_t(\tau)| \;<\; \infty$

Then the tracking failure probability is bounded as:

$$\mathbb{P}\Big[b^* \notin \mathcal{S}_t(\tau)\Big] + \mathbb{P}\left[\max_{b\in\mathcal{S}_t(\tau)\setminus\{b^*\}} s(I,t,b) \;\geq\; s(I,t,b^*) \,\middle|\, b^* \in \mathcal{S}_t(\tau)\right] \leq \underbrace{\eta(\tau)}_{\textit{ROI exclusion}} + \underbrace{\big(M(\tau)-1\big)\exp\left(-\frac{n\,\gamma(\tau)^2}{4\sigma^2}\right)}_{\textit{within-ROI mis-localization}}. \tag{12}$$

(See supplementary material for proof details.)

Theorem 2 reveals the bound of two failure modes under ROI-based search. Our MVLM confidence mechanism addresses both by dynamically switching between local ROI search (high confidence) and global re-localization (low confidence).

5. Experiments

5.1. Implementation details

For the visual encoder, we adopt HiViT [28] . For text encoder, we utilize CLIP-based text encoder [25]. For patch embedding, N_I, N_T, and C are set to 196, 77, and 512, respectively. To train our model, we utilize several visual-language tracking datasets containing image and text pairs including TNL2K [29], LaSOT [9], VastTrack [24]. To evaluate our model, we evaluate on TNL2K [29], LaSOT [9],

Table 1. Comparison of SOTA vision-language trackers on TNL2K, LaSOT, OTB99, and MGIT Benchmarks. The highest and second highest results are highlighted in **bold** and underline, respectively. † denotes the results obtained using the released official code.

Methodology	Year	Publication	TNL2K		LaSOT		OTB99		MGIT	
			PRE	AUC	PRE	AUC	PRE	AUC	PRE	AUC
Tracking-by-bbox and -language										
JointNLT [38]	2023	CVPR	58.1	56.9	63.6	60.4	85.6	65.3	44.5	61.0
QueryNLT [26]	2024	CVPR	58.7	57.8	62.5	54.2	88.2	66.7	-	-
MemVLT [10]	2024	NIPS	67.4	63.3	80.5	72.9	-	-	63.7	69.4
UVLTrack-B [23]	2024	AAAI	66.7	63.1	75.9	69.4	89.9	69.3	53.0†	63.7†
SUTrack-B224 [7]	2025	AAAI	67.9	65.0	80.5	**73.2**	93.4	70.8	-	-
MambaVLT [22]	2025	CVPR	69.9	66.5	71.0	66.6	**94.4**	**72.2**	58.9	69.9
DUTrack-256 [18]	2025	CVPR	70.6	64.9	**81.1**	73.0	93.9	70.9	65.4	70.0
MVLM (Ours)	-	-	**73.0**	**66.9**	80.4	72.7	94.0	70.6	**66.3**	**71.7**
Tracking-by-language										
TNL2K-1 [30]	2021	CVPR	6.0	11.0	49.0	51.0	24.0	19.0	-	-
CTRNLT [19]	2022	CVPRW	9.0	14.0	51.0	52.0	72.0	53.0	-	-
JointNLT [38]	2023	CVPR	55.0	54.6	59.3	56.9	77.6	59.2	43.8†	59.2†
QueryNLT [26]	2024	CVPR	53.0	53.3	55.0	54.2	81.0	**61.2**	-	-
UVLTrack-B [23]	2024	AAAI	57.2	55.7	61.0	57.2	79.1	60.1	44.6†	56.1†
MambaVLT [22]	2025	CVPR	58.9	**58.4**	57.2	55.8	79.2	58.9	50.3	**64.6**
MVLM (Ours)	-	-	**60.9**	57.8	**65.5**	**60.7**	**84.3**	60.7	**55.5**	63.5

Table 2. Ablation studies of component methods on TNL2K, LaSOT and OTB99.

	Global Search	Local Search	$\mathcal{L}_{CM}$	MVLM	TNL2K			LaSOT			OTB99		
					PRE	NPR	AUC	PRE	NPR	AUC	PRE	NPR	AUC
(A1)	✓				53.0	65.3	50.8	52.2	55.3	50.1	72.7	64.2	53.1
(A2)		✓			59.6	72.9	56.9	62.2	65.4	58.4	79.7	71.6	59.0
(A3)		✓	✓		60.1	73.3	57.2	65.3	68.1	60.6	82.8	74.0	59.8
(A4)	✓	✓	✓	✓	60.9	74.1	57.8	65.5	68.3	60.7	84.3	75.4	60.7

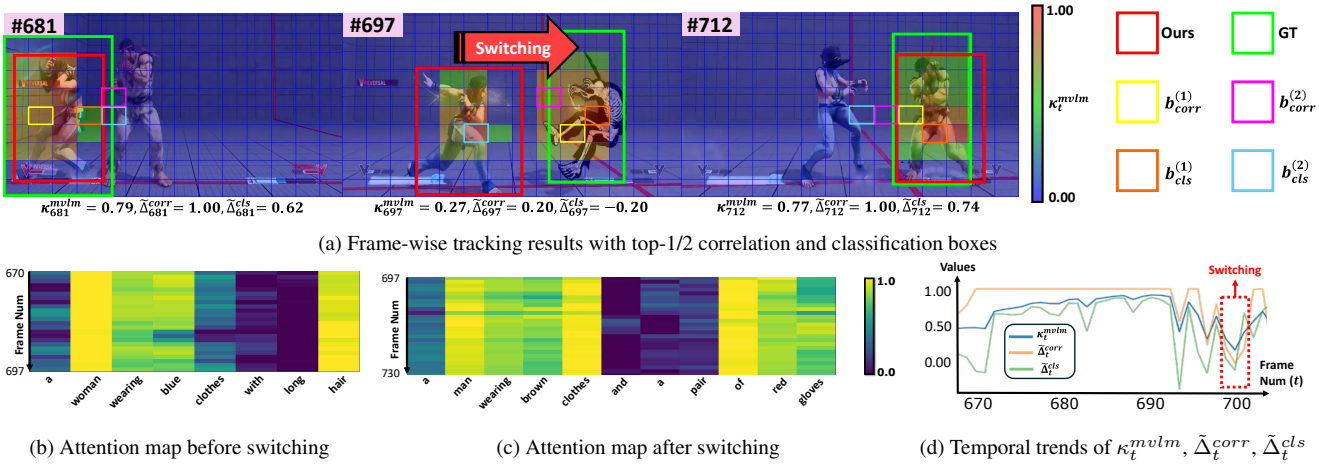
OTB99 [20] and MGIT [14]. We set the training epoch to 60 consisting of 100,000 with paired image and text, and learning rate to 0.0005. Batch size is set to 80. We also adopt AdamW optimizer. We train and evaluate our model using four NVIDIA RTX A6000 48GB GPUs. For all metrics for the accuracy on each set below, a higher score indicates better results.

5.2. Empirical validation of theoretical bounds

Mis-localization bound. To empirically verify the exponential decay structure predicted by Theorem 1, we select a representative frame from TNL2K [30] datasets and perform $K = 100$ independent stochastic forward passes by adding Gaussian noises ($\sigma = 0.1$) to simulate the token-level noise. For each candidate box b and stochastic pass k, we compute the alignment score $s^{(k)}(I, t, b) = \langle \bar{\mathbf{v}}^{(k)}(b), \mathbf{u}_t \rangle$ and estimate the empirical failure probability as $\hat{p}_{\text{fail}}(b) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}\{s^{(k)}(I, t, b) \geq s^{(k)}(I, t, b^*)\}$, where b^* is the GT box. We also compute the empirical correlation margin $\Delta\hat{\rho}(b) = \frac{1}{K}\sum_{k=1}^{K}[\rho^{(k)}(b^*) - \rho^{(k)}(b)]$, where $\rho^{(k)}(b) = \langle \bar{\boldsymbol{\mu}}^{(k)}(b), \mathbf{u}_t \rangle$ is the deterministic correlation in pass k. According to Theorem 1, the mis-localization probability decays exponentially as $p_{\text{fail}}(b) \lesssim \exp(-n_{\text{eff}} \cdot \Delta\rho(b)^2/2\sigma^2)$, where $n_{\text{eff}} = (|R(b^*)|^{-1} + |R(b)|^{-1})^{-1}$ is the effective sample size. Taking logarithms yields a linear relationship: $-\log p_{\text{fail}}(b) \approx c \cdot n_{\text{eff}} \cdot \Delta\rho(b)^2$ with $c = 1/(2\sigma^2)$. Figure 3-(a) presents a collapse plot for a representative frame, showing $y(b) = -\log \hat{p}_{\text{fail}}(b)$ versus $x(b) = n_{\text{eff}} \cdot (\Delta\hat{\rho}(b))^2$ across all candidate boxes within the frame. The plot reveals a clear linear relationship with fitted slope $\hat{c} = 263.6 \pm 0.356$ and high coefficient of determination $R^2 = 0.89$, corresponding to an estimated noise variance $\hat{\sigma}^2 = (2\hat{c})^{-1} \approx 0.0019$. We observe that the data points collapse onto a single line, confirming the universal exponential decay structure. This collapse plot directly validates our Correlation Margin (CM) loss design (Section 3.2): by maximizing $\Delta\rho(b)$ during training, we exponentially suppress mis-localization failures.

Tracking failure bound. We empirically validate the bound derived in Theorem 2. Our goal is to verify that the measured total failure probability $\hat{p}_{tot}(\tau)$ is always less than or equal to the calculated theoretical bound $\hat{B}_{tot}(\tau)$, i.e., $\hat{p}_{tot}(\tau) \leq \hat{B}_{tot}(\tau)$. We first define the empirical measures to validate the components of Theorem 2 as follows:

(1) $\hat{p}_{tot}(\tau)$: Empirical tracking failure rate, counting frames with $IoU(\hat{b}_t, b^*) < 0.5$.



(a) Frame-wise tracking results with top-1/2 correlation and classification boxes

(b) Attention map before switching

(c) Attention map after switching

(d) Temporal trends of κ_t^{mvlm}, $\tilde{\Delta}_t^{corr}$, $\tilde{\Delta}_t^{cls}$

Figure 4. The experiments of re-targeting the tracked object by changing natural languages. The natural language is updated from "a woman wearing blue clothes with long hair" to "a man wearing brown clothes and a pair of red gloves" at frame 697. (a) is the visualization of κ_t^{mvlm} scores, where higher values are highlighted with stronger red. (b) and (c) are visualization of attention distribution maps [11] before/after a language query update at frame t=697. (d) is represented to κ_t^{mvlm}, $\tilde{\Delta}_t^{\text{corr}}$ and $\tilde{\Delta}_t^{\text{cls}}$ the across frames. The red-dotted rectangle indicates the moments when the target identity is switched by the updated language description.

(2) $\hat{\eta}(\tau)$**:** Empirical ROI-exclusion rate, measuring the fraction of frames where $b^* \notin \mathcal{S}_t(\tau)$.

(3) $\hat{B}_{pre}(\tau)$**:** Within-ROI mis-localization bound computed as $\hat{B}_{pre}(\tau) = M(\tau)\exp(-\hat{c}_{pre} \cdot x_{pre}(\tau))$. Here, $x_{pre}(\tau) = n_{eff} \cdot \hat{\kappa}(\tau)$ where $\hat{\kappa}(\tau) = \text{median}_{b \in \mathcal{S}_t(\tau_t) \setminus \{b^*\}} (\Delta\hat{\rho}(b))^2$.

(4) $\hat{B}_{tot}(\tau)$**:** Theoretical bound computed as $\hat{B}_{tot}(\tau) = \hat{\eta}(\tau) + \hat{B}_{pre}(\tau)$, summing ROI-exclusion and within-ROI mis-localization terms.

We then evaluate these measures across all sequences in TNL2K and LaSOT datasets, varying τ from 0.1 to 0.9 to examine how the bound structure holds across different confidence thresholds Figure 3-(b) shows that $\hat{p}_{tot}$ (green) remains strictly below $\hat{B}_{tot}(\tau)$ (blue) across all τ values, confirming the inequality $\hat{p}_{tot} \leq \hat{B}_{tot}(\tau)$. This demonstrates that Theorem 2 provides a tight upper bound for predicting tracking failures and a valid guarantee on the worst-case failure rate under our MVLM confidence-based ROI filtering.

5.3. Comparison with state of the arts

In Table 1, we compare our tracker with state-of-the-art trackers on vision-language tracking benchmarks (TNL2K, LaSOT, OTB99 and MGIT).

Tracking-by-language: In this setting, we evaluate trackers where only language guidance is provided, without an initial bounding box. Across all benchmarks, our model achieves the best or second best performance. In particular, on TNL2K, LaSOT, OTB99 and MGIT, our tracker improves the precision (PRE) +2.0%, +4.5%, +3.3% and +5.2% compared to the second-best tracker, respectively. The PRE gain reflects the ability of the proposed method to stabilize localization in dynamic real-world conditions. This indicates that our approach supports the feasibility of language-guided tracking under template-free condition.

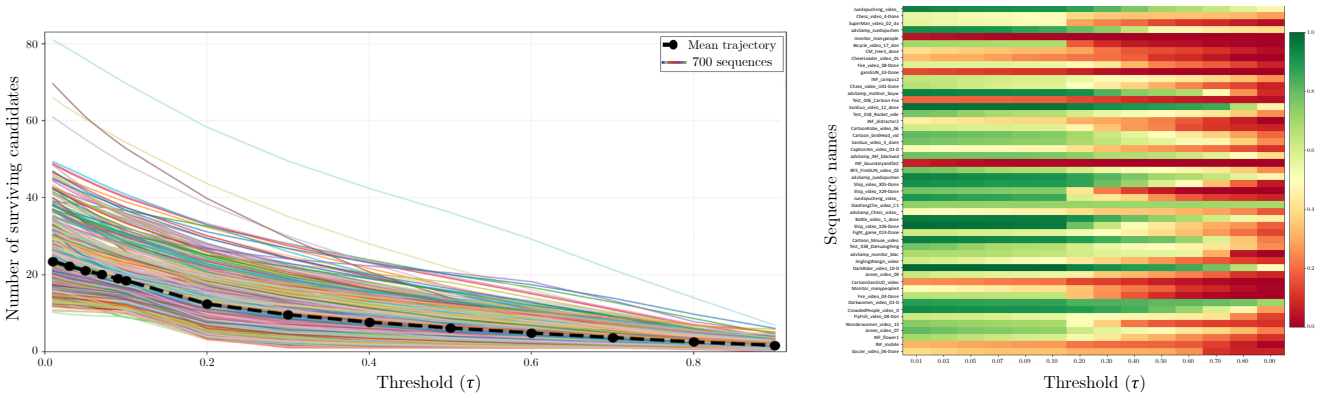
Tracking-by-bbox and -language: We further verify whether the proposed method can be extended not only to the template-free condition, but also to the template-based condition. To this end, we consider a tracking-by-bbox and -language paradigm, which is given both the language description and the visual template are provided. We implement this by combining the visual template with language. Specifically, we replace $P_t^* = [P_t^T; P_t^I]$ with $[P_t^T; P_t^I; P_t^R]$, where $P_t^R \in \mathbb{R}^{N_R \times C}$ is produced based on visual template and N_R is set to 49. We make no further changes aside from this. Table 1 shows that MVLM has the best PRE on TNL2K and MGIT. Although designed for template-free conditions, the method naturally extends to template-based tracking when visual templates are incorporated.

5.4. Ablation study

Table 2 shows the effect of our proposed components on TNL2K and OTB99, and LaSOT. To this end, we train and evaluate four tracker variants by altering the components.

- (A1) and (A2) training without $\mathcal{L}_{CM}$.
- (A3) training with $\mathcal{L}_{CM}$.
- (A4) training with $\mathcal{L}_{CM}$ and uses MVLM.

(A1) and (A2) select the final bounding box using $b_{cls}^{(1)}$. (A1) searches the entire frame, (A2) restricts the search to a local region around the predicted result via the local RoI search. When comparing (A1) with (A2), local ROI search improves the average PRE/AUC by 7.9%/6.8% across bench-



(a) Average RoI subset size vs. τ on TNL2K

(b) GT retention probability heatmap on TNL2K

Figure 5. (a) is the averaged RoI subset size $M(\tau)$ by MVLM, and black dotted line is the trajectory for averaged $M(\tau)$ across all sequences on TNL2K test-sets. (b) is GT retention probability heatmap on 50 sequences from TNL2K test set (for brevity). Each row denotes a sequence, and each column corresponds to τ.

marks. When comparing (A2) with (A3), $\mathcal{L}_{CM}$ improves the average PRE/AUC by 2.2%/1.1% across benchmarks, showing that the margin correlation signal improves target recognition. When comparing (A3) with (A4), $\mathcal{L}_{CM}$ and MVLM improves the average PRE/AUC by 0.8%/0.5% across benchmarks, showing that MVLM effectively selects reasonable ROI candidates.

5.5. Analysis of target re-assignment

A core challenge of template-free tracking is whether the tracker can shift to another target using only language. To explore this, we concatenate two TNL2K sequences that share the same background but contain different targets, and update the description from *a woman wearing blue clothes with long hair* to *a man wearing brown clothes and a pair of red gloves* at frame 697. As shown in Figure. 4-(b)/(c), the attention distribution is shifted from "woman/blue" to "man/brown/red gloves", and predicted bounding box smoothly switches to the new target in Figure. 4-(a). In Figure. 4-(d), κ_t^{mvlm}, $\tilde{\Delta}_t^{\text{corr}}$ and $\tilde{\Delta}_t^{\text{cls}}$ briefly drop temporarily at the switching point and then stabilize. This shows that the tracker recovers a consistent target identity under the new language description. The attention distribution is changed and the bounding box switch occurs together. This indicates that the language description defines the target identity for template-free tracking, and the target can be switched simply by modifying the description.

5.6. Analysis of GT retention

We analyze whether our MVLM mechanism effectively reduces distractors while retaining valid candidates. For each sequence in the TNL2K test set, we change τ and measure whether the subset $\mathcal{S}_t(\tau)$ contains the ground-truth box b^*. The total number of survived candidates $M(\tau)$ reaches up to 196, consisting of boxes generated from the tracking header. Figure 5–(a) shows that $M(\tau)$ decreases as τ increases. (e.g., it drops sharply from 196 to around 20 in average at $\tau = 0.01$). Figure 5–(b) visualizes the GT retention probability $\mathbb{P}\Big[\exists b \in \mathcal{S}_t(\tau) : \text{IoU}(b, b^*) > 0.5\Big]$. We find that most sequences consistently maintain a high GT retention probability at $\tau \leq 0.2$, even if our MVLM confidence-based mechanism filters out most of candidate boxes. Taken together Figure 5–(a) and Figure 5–(b), the tendency quantitatively demonstrates that the proposed confidence-based filtering suppresses low-confidence candidates while preserving GT retention stability. As a result, we confirm that the MVLM effectively prunes uncertain samples.

6. Conclusion

In this work, we address template-free tracking where the tracker localizes objects using only natural language, without visual templates or box initialization. We show that VL correlation margin is the fundamental tracking signal for reliable tracking: under a sub-Gaussian noise model, larger margins between the target and distractors exponentially reduce mis-localization probability. However, margins can degrade under challenging scenarios. To address this, we propose MVLM confidence that integrates correlation margins and temporal memory to dynamically gate between local ROI search and global re-localization. Theoretically, we derive bounds decomposing failure into ROI exclusion and within-ROI mis-localization, justifying MVLM as a theoretically grounded gating mechanism. Extensive experiments validate our theoretical predictions and achieve state-of-the-art performance under language-only guidance, demonstrating that margin-based correlation provides a reliable foundation for template-free tracking in challenging scenarios.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No.RS-2022-NR071978) and funded by the Ministry of Education (No.RS-2022-NR070869); supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No.RS-2022-II220448: Deep Total Recall, 10%); supported in part by INHA UNIVERSITY Research Grant.

References

[1] Seung-Hwan Bae. Deformable part region learning for object detection. In *AAAI*, pages 95–103, 2022. 2

[2] Seung-Hwan Bae. Deformable part region learning and feature aggregation tree representation for object detection. *IEEE TPAMI*, 45(9):10817–10834, 2023. 2

[3] Wenrui Cai, Qingjie Liu, and Yunhong Wang. Hiptrack: Visual tracking with historical prompts. In *CVPR*, pages 19258–19267, 2024. 1

[4] Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qiuhong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: A simplified architecture for visual object tracking. In *ECCV*, pages 375–392. Springer, 2022.

[5] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, pages 8126–8135, 2021.

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

[7] 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*, pages 2239–2247, 2025. 4, 6

[8] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *CVPR*, pages 13608–13618, 2022. 1, 2

[9] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, pages 5374–5383, 2019. 2, 5

[10] Xiaokun Feng, Xuchen Li, Shiyu Hu, Dailing Zhang, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. Memvlt: Vision-language tracking with adaptive memory-based prompts. In *NIPS*, 2024. 6

[11] Xiaokun Feng, Shiyu Hu, Xuchen Li, Dailing Zhang, Meiqi Wu, Jing Zhang, Xiaotang Chen, and Kaiqi Huang. Atctrack: Aligning target-context cues with dynamic target states for robust vision-language tracking. *ICCV*, 2025. 2, 7

[12] Zhihong Fu, Qingjie Liu, Zehua Fu, and Yunhong Wang. Stmtrack: Template-free visual tracking with space-time memory networks. In *CVPR*, pages 13774–13783, 2021. 2

[13] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *ECCV*, pages 146–164. Springer, 2022. 1

[14] Shiyu Hu, Dailing Zhang, Xiaokun Feng, Xuchen Li, Xin Zhao, Kaiqi Huang, et al. A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship. *NIPS*, 36:25007–25030, 2023. 2, 6

[15] Ben Kang, Xin Chen, Dong Wang, Houwen Peng, and Huchuan Lu. Exploring lightweight hierarchical vision transformers for efficient visual tracking. In *ICCV*, pages 9612–9621, 2023. 1

[16] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. 2

[17] Xin Li, Yuqing Huang, Zhenyu He, Yaowei Wang, Huchuan Lu, and Ming-Hsuan Yang. Citetracker: Correlating image and text for visual tracking. In *ICCV*, pages 9974–9983, 2023. 2

[18] Xiaohai Li, Bineng Zhong, Qihua Liang, Zhiyi Mo, Jian Nong, and Shuxiang Song. Dynamic updates for language adaptation in visual-language tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19165–19174, 2025. 2, 6

[19] Yihao Li, Jun Yu, Zhongpeng Cai, and Yuwen Pan. Cross-modal target retrieval for tracking by natural language. In *CVPRW*, pages 4931–4940, 2022. 6

[20] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6495–6503, 2017. 2, 6

[21] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, pages 38–55. Springer, 2024. 2

[22] Xinqi Liu, Li Zhou, Zikun Zhou, Jianqiu Chen, and Zhenyu He. Mambavlt: Time-evolving multimodal state space model for vision-language tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8731–8741, 2025. 6

[23] Yinchao Ma, Yuyang Tang, Wenfei Yang, Tianzhu Zhang, Jinpeng Zhang, and Mengxue Kang. Unifying visual and vision-language tracking via contrastive learning. In *AAAI*, pages 4107–4116, 2024. 1, 2, 6

[24] Liang Peng, Junyuan Gao, Xinran Liu, Weihong Li, Shaohua Dong, Zhipeng Zhang, Heng Fan, and Libo Zhang. Vasttrack: Vast category visual object tracking. *NIPS*, 37: 130797–130818, 2024. 5

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3, 5

[26] Yanyan Shao, Shuting He, Qi Ye, Yuchao Feng, Wenhan Luo, and Jiming Chen. Context-aware integration of language and visual references for natural language tracking. In *CVPR*, pages 19208–19217, 2024. 1, 2, 6

[27] Yuyang Tang, Yinchao Ma, and Tianzhu Zhang. Semantic-aware network for natural language tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2

[28] 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 TPAMI*, 2024. 3, 5

[29] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*, pages 1571–1580, 2021. 2, 5

[30] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *CVPR*, pages 13763–13773, 2021. 1, 2, 6

[31] Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *CVPR*, pages 9697–9706, 2023. 1

[32] Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, and Rongrong Ji. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *CVPR*, pages 19300–19309, 2024.

[33] Chenlong Xu, Bineng Zhong, Qihua Liang, Yaozong Zheng, Guorong Li, and Shuxiang Song. Less is more: Token context-aware learning for object tracking. In *AAAI*, pages 8824–8832, 2025.

[34] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *ICCV*, pages 10448–10457, 2021. 1

[35] Zhengyuan Yang, Tushar Kumar, Tianlang Chen, Jingsong Su, and Jiebo Luo. Grounding-tracking-integration. *IEEE TCSVT*, 31(9):3433–3443, 2020. 2

[36] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, pages 341–357. Springer, 2022. 1, 2, 4

[37] Xiaosong Zhang, Yunjie Tian, Lingxi Xie, Wei Huang, Qi Dai, Qixiang Ye, and Qi Tian. Hivit: A simpler and more efficient design of hierarchical vision transformer. *ICLR*, 2023. 2

[38] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *CVPR*, pages 23151–23160, 2023. 1, 2, 6