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

Supplementary Material

In this supplementary material, we provide a more detailed analysis of our method and comparison results.

- In Section 1, we provide detailed proofs for Section 3 and Section 4 of our main paper.
- In Section 2, we provide more details on our tracking header network and losses.
- In Section 3, we provide more detailed description of global and local ROI search.
- In Section 4, we provide sensitivity analysis for τ, $\psi_{out}, \alpha_{corr}$, λ, and ω.
- In Section 5, we provide additional qualitative and quantitative experimental results and analyses.
- In Section 6, we provide analysis of computational cost and tracking speed.

1. Theoretical Analysis and Proofs

This section provides detailed mathematical justifications for the theoretical results presented in Section 4 of our main our main paper. We first establish the necessary notation and assumptions, then present complete proofs for Theorems 1 and 2.

1.1. Preliminaries and Notation

For completeness, we restate the key definitions and assumptions used in our theoretical analysis.

Definition 1 (Visual search token)**.** *Let $F \in \mathbb{R}^{H\times W\times d}$ be a feature map of one frame. Define the spatial index set*

$$\Omega \;=\; \{(i,j) \,:\, 1 \le i \le H,\; 1 \le j \le W\}.$$

For $\mathbf{x} \in \Omega$, the vision token embedding is $\mathbf{v}_{\mathbf{x}} \equiv \mathbf{v}_{(i,j)} \in \mathbb{R}^d$, where the subscript $\mathbf{x} = (i,j)$ denotes a vision token location within the image grid.

Let $\mathbf{v}_{t,\mathbf{x}}$ denote the vision token at location $\mathbf{x}$ at frame t and $\mathbf{u}_t$ the language embedding given at frame t. For brevity, we omit the frame index t for visual tokens when the context is clear.

Definition 2 (Region score via vision-language correlation)**.** *A candidate box b corresponds to a token index set $R(b) \subset \Omega$ (tokens whose cell centers fall inside b). We denote b^* as the ground-truth bounding box of the target and $R^* = R(b^*)$ as the set of token indices contained in b^*.*

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

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

where $\bar{\mathbf{v}}_t(b) = \frac{1}{|R(b)|}\sum_{\mathbf{x}\in R(b)} \mathbf{v}_{\mathbf{x}}$ is the mean vision embedding over region $R(b)$.

Assumption 1.1 (Basic normalization and sub-Gaussian noise model)**.** *We assume $\|\mathbf{u}_t\|_2 = 1$ and $\|\mathbf{v}_{\mathbf{x}}\|_2 \le 1$ for all $\mathbf{x} \in \Omega$. Write $\mathbf{v}_{\mathbf{x}} = \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\varepsilon}_{\mathbf{x}}$, where $\boldsymbol{\mu}_{\mathbf{x}}$ represents the signal component and $\boldsymbol{\varepsilon}_{\mathbf{x}}$ is zero-mean noise.*

We assume $\langle \boldsymbol{\varepsilon}_{\mathbf{x}},\, \mathbf{u}_t\rangle$ is sub-Gaussian with proxy variance σ^2, meaning for all $\lambda \in \mathbb{R}$:

$$\mathbb{E}\big[e^{\lambda\langle \boldsymbol{\varepsilon}_{\mathbf{x}},\, \mathbf{u}_t\rangle}\big] \le e^{\lambda^2\sigma^2/2}.$$

Within a box, tokens are either independent or exhibit sufficiently weak dependence such that averages over $R(b)$ concentrate at rate $O(1/|R(b)|)$. Additionally, disjoint regions $R(b^*)$ and $R(b)$ are assumed independent.

Definition 3 (Region mean correlation)**.** *For a box b with region $R(b)$, define the region-mean signal vector*

$$\bar{\boldsymbol{\mu}}(b) = \frac{1}{|R(b)|}\sum_{\mathbf{x}\in R(b)} \boldsymbol{\mu}_{\mathbf{x}}, \quad (2)$$

and its alignment with the language embedding,

$$\rho(b) = \langle \bar{\boldsymbol{\mu}}(b),\, \mathbf{u}_t\rangle. \quad (3)$$

We call $\rho(b)$ the noiseless correlation *of box b with the language query.*

Assumption 1.2 (Correlation margin)**.** *There exists $\gamma > 0$ such that for all distractor boxes $b \ne b^*$,*

$$\rho(b^*) - \rho(b) \;\ge\; \gamma. \quad (4)$$

Equivalently, the ground-truth region maintains strictly higher noiseless correlation with the language description than any competing candidate.

Remark 1.1. *Assumption 1.2 formalizes the intuition that a semantically precise language description combined with visually distinctive target appearance ensures sufficient separation in the correlation space. In practice, γ depends on both the specificity of the language query and the visual distinctiveness of the target relative to background distractors.*

1.2. Proof of Theorem 1: Mis-localization Bound

We now present the complete proof of Theorem 1 from our main paper, which establishes an exponential bound on mis-localization probability under sub-Gaussian noise.

Theorem 1 (Mis-localization bound under sub-Gaussian noise)**.** *Under Assumptions 1.1 and 1.2, with $n^* = |R^*|$ and $n_b = |R(b)|$, the probability that any distractor box b achieves a higher score than the ground-truth box b^* is bounded as:*

$$\mathbb{P}\Big[\max_{b\neq b^*} s(I,t,b) \;\geq\; s(I,t,b^*)\Big] \leq \sum_{b\neq b^*} \exp\left(-\frac{\big(\rho(b^*)-\rho(b)\big)^2}{2\sigma^2\Big(\frac{1}{n^*}+\frac{1}{n_b}\Big)}\right). \tag{5}$$

Proof. The proof proceeds in three steps: (1) decompose the score gap into signal and noise, (2) bound the tail probability for a single distractor, and (3) apply the union bound.

Step 1: Score gap decomposition. Consider any distractor $b \neq b^*$. The random score gap is:

$$\begin{aligned} s(I,t,b^*) - s(I,t,b) &= \underbrace{\rho(b^*)-\rho(b)}_{\text{signal: }\geq\gamma} \\ &+ \underbrace{\frac{1}{n^*}\sum_{\mathbf{x}\in R^*}\langle \boldsymbol{\varepsilon}_{\mathbf{x}},\, \mathbf{u}_t\rangle - \frac{1}{n_b}\sum_{\mathbf{x}\in R(b)}\langle \boldsymbol{\varepsilon}_{\mathbf{x}},\, \mathbf{u}_t\rangle}_{\text{noise: }=Z_b}. \end{aligned} \tag{6}$$

The first term is the deterministic margin guaranteed by Assumption 1.2. The second term Z_b represents the combined noise from both regions.

Step 2: Sub-Gaussian concentration. By Assumption 1.1, each $\langle \boldsymbol{\varepsilon}_{\mathbf{x}},\, \mathbf{u}_t\rangle$ is sub-Gaussian with proxy variance σ^2. Since the two regions R^* and $R(b)$ are disjoint and independent, the difference of averages Z_b is also sub-Gaussian.

The proxy variance of Z_b is:

$$\mathrm{Var}_{\text{proxy}}(Z_b) = \frac{\sigma^2}{n^*} + \frac{\sigma^2}{n_b} = \sigma^2\Big(\frac{1}{n^*}+\frac{1}{n_b}\Big). \tag{7}$$

For a sub-Gaussian random variable Z with zero mean and proxy variance v^2, the tail bound states:

$$\mathbb{P}[Z \leq -\delta] \leq \exp\Big(-\frac{\delta^2}{2v^2}\Big), \quad \forall \delta > 0.$$

Applying this to our setting, the event $\{s(I,t,b) \geq s(I,t,b^*)\}$ is equivalent to $\{s(I,t,b^*) - s(I,t,b) \leq 0\}$, which means $\{\rho(b^*)-\rho(b)+Z_b \leq 0\}$ or $\{Z_b \leq -(\rho(b^*)-\rho(b))\}$.

Since $\rho(b^*)-\rho(b) \geq \gamma > 0$, we have:

$$\begin{aligned} \mathbb{P}\big[s(I,t,b) \geq s(I,t,b^*)\big] &= \mathbb{P}\big[Z_b \leq -(\rho(b^*)-\rho(b))\big] \\ &\leq \exp\left(-\frac{(\rho(b^*)-\rho(b))^2}{2\sigma^2\big(\frac{1}{n^*}+\frac{1}{n_b}\big)}\right). \end{aligned} \tag{8}$$

Step 3: Union bound over all distractors. Let $\mathcal{B}_t$ denote the set of all candidate boxes at frame t. The event that the maximum score over all distractors exceeds the ground-truth score is:

$$\Big\{\max_{b\neq b^*} s(I,t,b) \geq s(I,t,b^*)\Big\} = \bigcup_{b\neq b^*}\big\{s(I,t,b) \geq s(I,t,b^*)\big\}.$$

By the union bound:

$$\mathbb{P}\Big[\max_{b\neq b^*} s(I,t,b) \geq s(I,t,b^*)\Big] \leq \sum_{b\neq b^*}\mathbb{P}\big[s(I,t,b) \geq s(I,t,b^*)\big].$$

Substituting the bound from equation (8) yields the desired result (5). □

Remark 1.2 (Interpretation)**.** *Theorem 1 reveals that the mis-localization probability decays* exponentially *with the squared correlation margin $(\rho(b^*)-\rho(b))^2$. This exponential dependence justifies our Correlation Margin (CM) loss design (Section 3.2 of the main paper), which directly maximizes the margin $\Delta\rho(b) = \rho(b^*)-\rho(b)$ during training.*

The bound also shows that larger regions (larger n^* and n_b) provide better concentration by averaging out noise, while higher noise variance σ^2 increases the failure probability. The factor $(\frac{1}{n^*}+\frac{1}{n_b})$ represents the effective inverse sample size.

1.3. Proof of Theorem 2: MVLM Re-localization Bound

While Theorem 1 establishes that sufficient correlation margin ensures reliable localization, in practice distractors or appearance shifts can reduce this margin. Our MVLM confidence mechanism addresses this by dynamically selecting a Region of Interest (ROI) through confidence-based filtering. However, this introduces two potential failure modes: (i) the ground-truth box may be excluded from the ROI (ROI-exclusion), and (ii) within the ROI, a distractor may still achieve higher score than the ground-truth (within-ROI mis-localization).

Theorem 2 formalizes the overall failure probability under ROI-based search.

Theorem 2 (MVLM confidence re-localization bound)**.** *Under the sub-Gaussian noise model (Assumption 1.1) with a common token-level proxy variance σ^2, assume:*

(a) Equal region size: $|R(b)| = n$ *for all* $b \in \mathcal{S}_t(\tau)$

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

(c) Minimum margin within ROI: $\rho(b^*) - \rho(b) \;\ge\; \gamma(\tau)$ *for all* $b \in \mathcal{S}_t(\tau) \setminus \{b^*\}$

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

where $\mathcal{S}_t(\tau) = \{b \in \mathcal{B}_t : \kappa_t^{\mathrm{mvlm}}(b) \ge \tau\}$ *is the ROI selected by thresholding the MVLM confidence at level* τ.

Then the overall tracking failure probability is bounded as:

$$\begin{aligned}
&\mathbb{P}\Big[\textit{tracking fails at frame } t\Big] \\
&= \mathbb{P}\Big[b^* \notin \mathcal{S}_t(\tau)\Big] \\
&\quad + \mathbb{P}\left[\max_{b \in \mathcal{S}_t(\tau)\setminus\{b^*\}} s(I,t,b) \ge s(I,t,b^*) \,\middle|\, b^* \in \mathcal{S}_t(\tau)\right] \\
&\le \underbrace{\eta(\tau)}_{\textit{ROI exclusion}} + \underbrace{\big(M(\tau)-1\big)\exp\!\left(-\frac{n\,\gamma(\tau)^2}{4\sigma^2}\right)}_{\textit{within-ROI mis-localization}}.
\end{aligned} \tag{9}$$

Proof. The proof decomposes the failure probability into two disjoint events and bounds each separately.

Step 1: Event decomposition. Define the following events:

- $E_{\mathrm{out}} := \{b^* \notin \mathcal{S}_t(\tau)\}$ (ROI exclusion: ground-truth is filtered out)
- $E_{\mathrm{in}} := \{\exists\, b \in \mathcal{S}_t(\tau) \setminus \{b^*\} : s(I,t,b) \ge s(I,t,b^*)\}$ (within-ROI mis-localization given $b^* \in \mathcal{S}_t(\tau)$)

The overall tracking failure occurs if either the ground-truth is excluded from the ROI or a distractor within the ROI achieves higher score:

$$\{\text{tracking fails}\} = E_{\mathrm{out}} \cup (E_{\mathrm{in}} \cap E_{\mathrm{out}}^c).$$

By the law of total probability:

$$\begin{aligned}
\mathbb{P}(\text{fail}) &= \mathbb{P}(E_{\mathrm{out}}) + \mathbb{P}(E_{\mathrm{in}} \cap E_{\mathrm{out}}^c) \\
&= \mathbb{P}(E_{\mathrm{out}}) + \mathbb{P}(E_{\mathrm{in}} \mid E_{\mathrm{out}}^c)\cdot\mathbb{P}(E_{\mathrm{out}}^c) \\
&\le \mathbb{P}(E_{\mathrm{out}}) + \mathbb{P}(E_{\mathrm{in}} \mid b^* \in \mathcal{S}_t(\tau)).
\end{aligned} \tag{10}$$

Step 2: Bounding ROI exclusion probability. By assumption (b), we directly have:

$$\mathbb{P}(E_{\mathrm{out}}) = \mathbb{P}\big[b^* \notin \mathcal{S}_t(\tau)\big] \le \eta(\tau).$$

The quantity $\eta(\tau)$ captures how conservative the MVLM confidence thresholding is. A well-designed confidence measure should yield small $\eta(\tau)$ at operating thresholds, ensuring the ground-truth is retained in the ROI with high probability.

Step 3: Bounding within-ROI mis-localization. Conditioning on $b^* \in \mathcal{S}_t(\tau)$, we need to bound the probability that some distractor in the ROI achieves higher score than b^*.

For any fixed $b \in \mathcal{S}_t(\tau) \setminus \{b^*\}$, define the score gap:

$$\Delta_b := s(I,t,b^*) - s(I,t,b) = \big(\rho(b^*) - \rho(b)\big) + Z_b,$$

where Z_b is the noise term as in equation (6).

By assumption (a), all boxes in the ROI have equal region size n. Therefore, by the sub-Gaussian property:

$$Z_b \sim \text{sub-Gaussian with proxy variance } \sigma^2\Big(\frac{1}{n}+\frac{1}{n}\Big) = \frac{2\sigma^2}{n}.$$

By assumption (c), the minimum margin within the ROI is $\gamma(\tau)$:

$$\rho(b^*) - \rho(b) \ge \gamma(\tau) \quad \forall b \in \mathcal{S}_t(\tau) \setminus \{b^*\}.$$

Using the sub-Gaussian tail bound:

$$\begin{aligned}
\mathbb{P}\big[s(I,t,b) \ge s(I,t,b^*) \mid b^* \in \mathcal{S}_t(\tau)\big] &= \mathbb{P}[\Delta_b \le 0] \\
&= \mathbb{P}\big[Z_b \le -(\rho(b^*) - \rho(b))\big] \\
&\le \mathbb{P}\big[Z_b \le -\gamma(\tau)\big] \\
&\le \exp\left(-\frac{\gamma(\tau)^2}{2\cdot\frac{2\sigma^2}{n}}\right) \\
&= \exp\left(-\frac{n\,\gamma(\tau)^2}{4\sigma^2}\right).
\end{aligned} \tag{11}$$

Step 4: Union bound over ROI candidates. By assumption (d), the ROI contains at most $M(\tau)$ candidates, so there are at most $M(\tau)-1$ distractors. Applying the union bound:

$$\begin{aligned}
&\mathbb{P}\left[\max_{b \in \mathcal{S}_t(\tau)\setminus\{b^*\}} s(I,t,b) \ge s(I,t,b^*) \,\middle|\, b^* \in \mathcal{S}_t(\tau)\right] \\
&\le \sum_{b \in \mathcal{S}_t(\tau)\setminus\{b^*\}} \mathbb{P}\big[s(I,t,b) \ge s(I,t,b^*) \mid b^* \in \mathcal{S}_t(\tau)\big] \\
&\le \big(M(\tau)-1\big)\exp\left(-\frac{n\,\gamma(\tau)^2}{4\sigma^2}\right).
\end{aligned} \tag{12}$$

Step 5: Combining the bounds. Substituting equations (12) and the ROI exclusion bound into equation (10) yields the final result (9). □

Remark 1.3 (Practical implications)**.** *Theorem 2 reveals the fundamental trade-off in ROI-based search:*

- ***Increasing*** τ *makes the ROI more selective:* $M(\tau)$ *decreases (fewer distractors) and* $\gamma(\tau)$ *typically increases (higher quality survivors), reducing within-ROI mis-localization. However,* $\eta(\tau)$ *increases (higher chance of excluding* b^**), raising the ROI exclusion risk.*
- ***Decreasing*** τ *admits more candidates:* $M(\tau)$ *increases and* $\gamma(\tau)$ *decreases, making within-ROI mis-localization more likely, but* $\eta(\tau)$ *decreases, reducing ROI exclusion.*

Our MVLM confidence mechanism dynamically adapts τ *(or equivalently, switches between local ROI search and global re-localization) to balance these two failure modes based on the current tracking state.*

Corollary 1 (Sufficient design rule)**.** *Fix a target failure probability* $\delta \in (0, 1)$*. If the operating threshold* τ *satisfies*

$$\eta(\tau) + \big(M(\tau) - 1\big) \exp\left(- \frac{n\,\gamma(\tau)^2}{4\sigma^2} \right) \leq \delta, \tag{13}$$

then the tracking failure probability at frame t *is at most* δ*.*

This provides a principled guideline for selecting τ *given estimates of* $\eta(\tau)$*,* $M(\tau)$*, and* $\gamma(\tau)$ *from validation data or online statistics.*

1.4. Robustness to Non-Gaussian Noise

Our theoretical analysis models score-level uncertainty with a zero-mean sub-Gaussian noise to derive clean bounds via MGF-based concentration. This assumption aligns well with typical tracking variability, such as mild appearance or illumination changes, moderate blur, and feature noise. To assess the robustness of our framework beyond this regime, we replace the noise term with Laplace and Gamma perturbations. Intuitively, Laplace noise reflects heavier-tailed deviations, such as brief occlusions, abrupt motion or defocus, and transient strong distractors. On the other hand, Gamma noise captures asymmetric score bursts, which can occur as spuriously high correlation peaks from textured backgrounds. As shown in Figure 1, while the sub-Gaussian setting best matches our theoretical assumptions, MVLM remains highly effective under these non-Gaussian settings. This empirical evidence suggests that our memory-gated decision mechanism and correlation margin are not strictly tied to a single noise family, demonstrating practical robustness in diverse and challenging tracking conditions.

2. Template-free Tracking Framework Details

2.1. Tracking Head Architecture

We adopt heatmap-based prediction heads consisting of three parallel branches: classification, offset regression, and size regression. Each head is composed of a stack of four

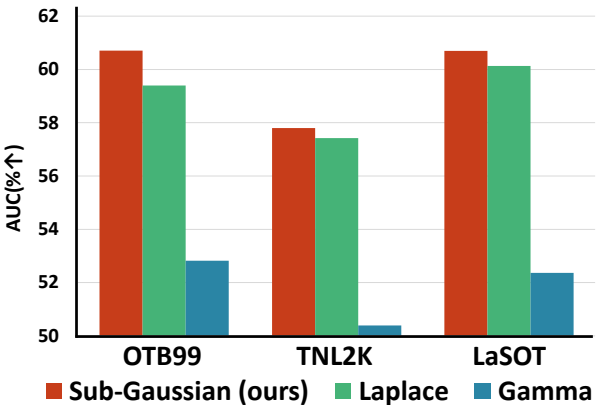


Figure 1. Robustness evaluation against non-Gaussian noise.

Conv-BN-ReLU blocks followed by a task-specific 1×1 convolution. Their outputs are formulated as:

$$\begin{aligned} F_{\text{cls}} &\in \mathbb{R}^{1\times\sqrt{N_I}\times\sqrt{N_I}}, \\ F_{\text{off}} &\in \mathbb{R}^{2\times\sqrt{N_I}\times\sqrt{N_I}}, \\ F_{\text{size}} &\in \mathbb{R}^{2\times\sqrt{N_I}\times\sqrt{N_I}}. \end{aligned} \tag{14}$$

During training and inference, we utilize an argmax-based decoding rule. The position (x_m, y_m) is determined by the peak of the classification heatmap:

$$(x_m, y_m) = \underset{(x,y)}{\operatorname{argmax}}\, F_{\text{cls}}(x, y). \tag{15}$$

We then extract the corresponding offset and size values as:

$$(d_x, d_y) = F_{\text{off}}(x_m, y_m), \quad (d_w, d_h) = F_{\text{size}}(x_m, y_m). \tag{16}$$

Here, (d_x, d_y) represent the offsets, indicating how far the true object center deviates from (x_m, y_m), and (d_w, d_h) denote the normalized width and height. By combining these values, the final bounding box b is reconstructed as:

$$b = (\hat{x}, \hat{y}, \hat{w}, \hat{h}) = (x_m + d_x,\, y_m + d_y,\, d_w,\, d_h). \tag{17}$$

2.2. Detailed Loss Formulation

We employ a composite loss function $\mathcal{L}_{\text{total}}$ to train the framework end-to-end. Given the ground-truth box $b_g = (x_g, y_g, w_g, h_g)$, the classification branch is supervised with a weighted focal loss [2] $\mathcal{L}_{\text{cls}}(F_{\text{cls}}, G_{\text{cls}})$ using a Gaussian heatmap G_{cls} centered at (x_g, y_g). Box quality is refined through a GIoU loss $\mathcal{L}_{\text{GIoU}}$ [4] and L1 penalty $\mathcal{L}_{\text{L1}}$ on the peak coordinates. To sum up, these losses are linearly combined and called track loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}}(F_{\text{cls}}, G_{\text{cls}}) + \alpha_{\text{GIoU}} \cdot \mathcal{L}_{\text{GIoU}}(b, b_g) + \alpha_{\text{L1}} \cdot \mathcal{L}_{\text{L1}}((x_p, y_p), (x_g, y_g)) + \mathcal{L}_{\text{CM}}$. Therefore, the total loss $\mathcal{L}_{\text{total}}$ is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{track}} + \mathcal{L}_{\text{CM}}, \tag{18}$$

where α_{GIoU} and α_{L1} are set to 2.0 and 5.0, respectively. The correlation margin loss $\mathcal{L}_{CM}$ is defined in Section 3.2 of our main paper.

3. Methodology for Scaling Search Region

We present the implementation details for the *Local search* and *Global search* strategies used to derive the visual search region.

3.1. Local ROI Search

Local search restricts the visual search region to a square region centered on the target predicted at the previous frame. Given the predicted bounding box $\hat{b}_{t-1} = (x_{t-1}, y_{t-1}, w_{t-1}, h_{t-1})$ from frame $t-1$, we define the geometric center $(c_x, c_y) = (x_{t-1} + w_{t-1}/2, y_{t-1} + h_{t-1}/2)$. The search region side length s_{local} is determined by the geometric mean of the target dimensions scaled by a search factor f_{search} (typically set to 4.0) to include sufficient surrounding context:

$$s_{local} = \lceil \sqrt{w_{t-1} \cdot h_{t-1}} \times f_{search} \rceil \tag{19}$$

We extract a square crop of size $s_{local} \times s_{local}$ from I_t centered at (c_x, c_y). If this crop extends beyond the image borders, zero-padding is applied. The cropped region is then resized to the fixed network input resolution $S_{in} \times S_{in}$ (e.g., 224×224). The corresponding scaling factor is defined as $r = S_{in}/s_{local}$, and let (x', y', w', h') denote the predicted bounding box in the local search region. Mapping back to the original image space I_t yields:

$$\hat{b}_t = \left(\frac{x'}{r} + c_x - \frac{s_{local}}{2}, \quad \frac{y'}{r} + c_y - \frac{s_{local}}{2}, \quad \frac{w'}{r}, \quad \frac{h'}{r} \right) \tag{20}$$

As mentioned in Section 1 on our main paper, local ROI search is a preprocessing strategy widely adopted by many existing trackers. It operates under the assumption that the position of tracked object does not change significantly from the previous frame. It removes similar appearance objects located elsewhere in the frame. However, this approach requires a prior tracking result to be available, making it unsuitable for initializing template-free tracking. We replace this mechanism with MVLM.

3.2. Global ROI Search

Global search processes the entire input frame I_t to localize the target object. Given the input frame $I_t \in \mathbb{R}^{H \times W}$, it is resized to the same square resolution $S_{\text{in}} \times S_{\text{in}}$ as in local search. This induces anisotropic scaling with horizontal and vertical factors:

$$r_x = \frac{S_{in}}{W}, \quad r_y = \frac{S_{in}}{H} \tag{21}$$

The visual encoder thus sees the full field of view without any spatial cropping. Let (x', y', w', h') denote the predicted box in this global search region. We map it back to the original resolution by inverting the scales:

$$\hat{b}_t = (x'/r_x, \quad y'/r_y, \quad w'/r_x, \quad h'/r_y) \tag{22}$$

4. Sensitivity Analysis for hyper-parameters

In this section, we perform a detailed sensitivity analysis of the key hyperparameters used in the MVLM confidence mechanism and the memory-gated tracking strategy. To ensure robustness, we conduct these ablations on the OTB99 [1] dataset. Based on the results, we adopt the following default settings for all experiments: confidence threshold $\tau = 0.1$, spatial exclusion threshold $\psi_{out} = 0.0$, blending weight $\alpha_{corr} = 0.1$, Forgetting factor $\lambda = 0.4$, and memory factor $\omega = 0.4$.

4.1. Analysis for Confidence Threshold τ

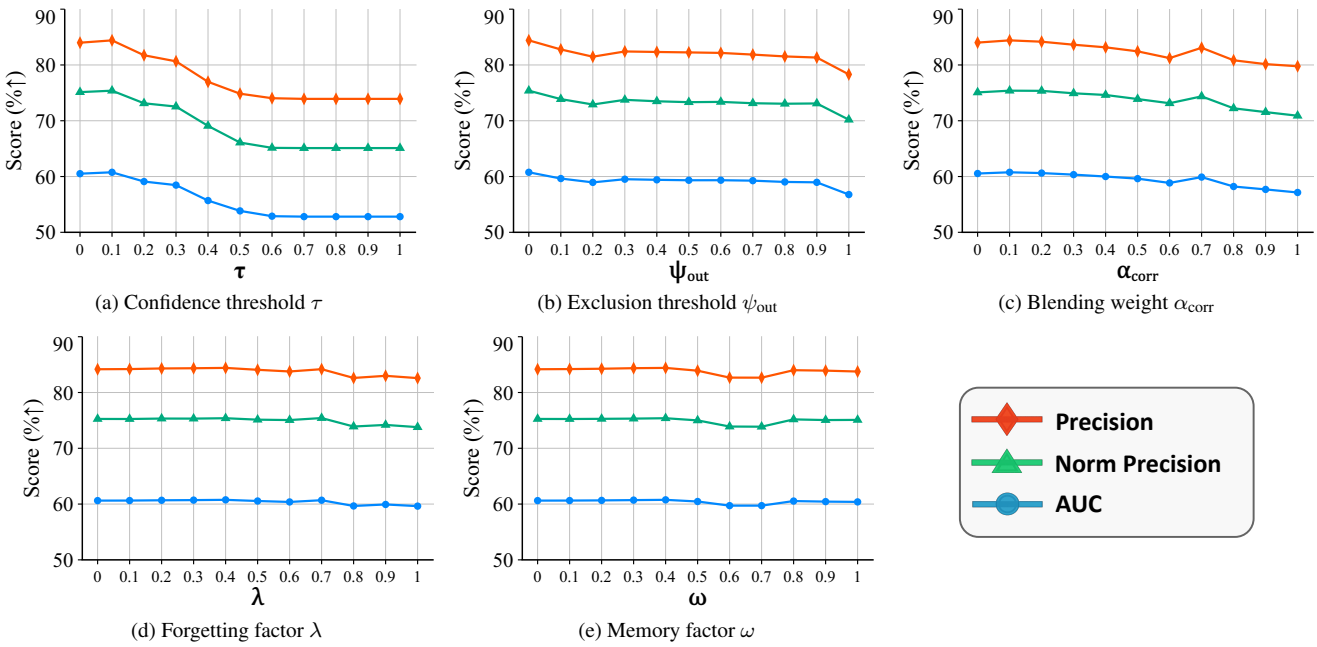
The threshold τ serves as the gating parameter that determines the size of the search region set $\mathcal{S}_t(\tau)$ defined in Eq. (8) of the main paper. It controls the trade-off between local ROI search and global re-localization. As shown in Figure 2-(a), we observe the following behaviors: The tracker retains high Precision and AUC when τ is low ($\tau \leq 0.1$). This suggests that allowing a slightly larger number of candidates into the ROI is safer than aggressive pruning, as it minimizes the ROI exclusion probability ($\eta(\tau)$). As τ increases beyond 0.2 ($\tau > 0.2$), performance drops sharply (e.g., AUC drops from 60.76% to 52.81%). A high threshold makes the ROI selection overly conservative, frequently excluding the ground truth during appearance changes and triggering potentially unstable global relocalizations. Therefore, we set $\tau = 0.1$ to balance suppressing distractors while retaining the target.

4.2. Analysis for Exclusion Threshold ψ_{out}

The exclusion threshold ψ_{out} defines the outside candidate pool $\mathcal{B}_t^{out}$, as detailed in Section 3.3 of the main paper. By enforcing that candidates satisfy $\text{IoU}(b', b) \leq \psi_{out}$, we ensure that the margin is computed against a geometrically distinct distractor rather than a highly overlapping candidate. Figure 2-(b) demonstrates that the framework achieves the best performance at $\psi_{out} = 0.0$, with a gradual decline as ψ_{out} increases. This indicates that computing the margin against fully non-overlapping distractors yields the most reliable separation signal between the target and background candidates. We therefore set $\psi_{out} = 0.0$.

4.3. Analysis for Blending Weight α_{corr}

The MVLM confidence κ_t^{vlm} is a weighted combination of the vision-language correlation margin $\tilde{\Delta}_t^{corr}$ and the classification margin $\tilde{\Delta}_t^{cls}$, controlled by $\kappa_t^{vlm} := \alpha_{corr} \tilde{\Delta}_t^{corr} +$



(a) Confidence threshold τ (b) Exclusion threshold ψ_{out} (c) Blending weight α_{corr}

(d) Forgetting factor λ (e) Memory factor ω

Figure 2. Sensitivity analysis of MVLM hyperparameters on OTB99. Each plot varies one parameter while fixing the others at default values (τ=0.1, Ψ_{out}=0.0, α_{corr}=0.1, λ=0.4, ω=0.4).

$\alpha_{cls}\tilde{\Delta}_t^{cls}$ subject to the constraint $\alpha_{corr} + \alpha_{cls} = 1$. Note that these two weights are coupled: setting $\alpha_{corr} = 0.0$ automatically sets $\alpha_{cls} = 1.0$, and as α_{corr} increases, the influence of α_{cls} decreases proportionally. Therefore, the x-axis in Figure 2-(c) only shows α_{corr} as $\alpha_{cls} = 1 - \alpha_{corr}$. Figure 2-(c) shows that a balanced fusion ($\alpha_{corr} = 0.1$) yields the most robust tracking performance in terms of all the three evaluation scores. This indicates that the classification score carries the dominant discriminative power for reliable target recognition, while the VL correlation margin acts as a lightweight complementary signal that provides the accuracy gain. Therefore, we set $\alpha_{corr} = 0.1$.

4.4. Analysis for Memory Factors λ and ω

In addition to the gating and thresholding parameters, we evaluate the sensitivity of our memory mechanism controlled by the exponentially weighted moving average (EWMA) factors, λ and ω. We optimize these parameters using a held-out validation split once and keep them fixed across all benchmarks to ensure generalization. As illustrated in Figure 2-(d), the tracking performance remains very stable over a wide range of $\lambda \in [0.0, 0.7]$, with the optimum at $\lambda = 0.4$. A moderate λ can balance responsiveness to abrupt confidence changes with temporal noise suppression. Figure 2-(e) shows that ω exhibits the lowest sensitivity among all hyperparameters, with AUC varying by only 0.37% across the entire range $\omega \in [0.0, 1.0]$. These results indicate that MVLM is not highly sensitive to the exact choices of these memory factors, demonstrating the robustness of our temporal smoothing design.

(a) VL correlation map (b) κ_t^{mvlm} map (c) Tracking result & GT

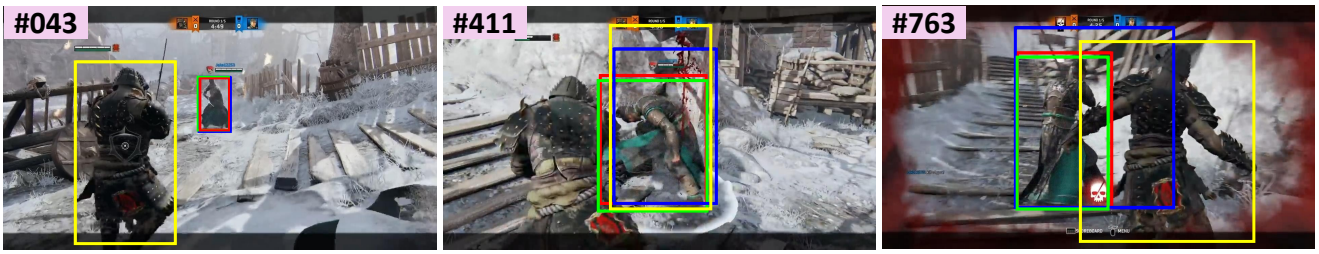
Figure 3. Visualization of robustness to implicit query ("the man in yellow"). Red/Green box are prediction/GT boxes, respectively.

5. Additional Experimental Results

In this section, we provide additional qualitative and quantitative results that complement the analyses presented in our main paper. These experiments aim to more clearly illustrate how MVLM behaves under diverse tracking conditions and how it compares against existing vision-language trackers.

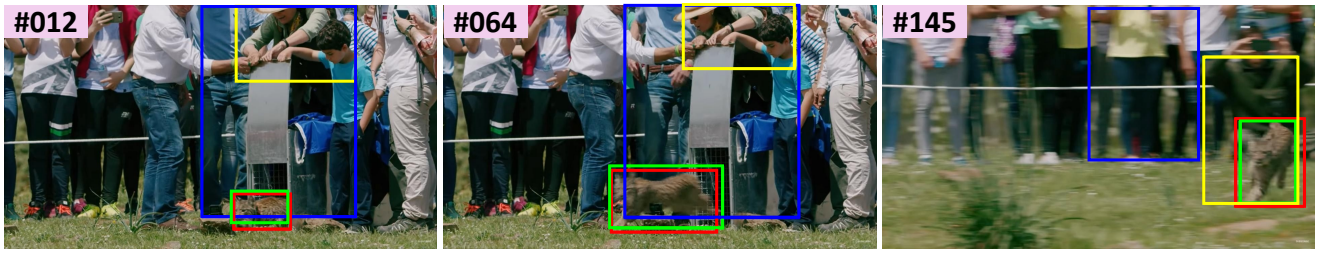
5.1. Tracking under Implicit/Ambiguous Queries

As illustrated in Figure 3, implicit or high-level semantic queries (e.g., "the man in yellow") can yield multiple activated peaks in the vision-language (VL) correlation map, as multiple instances may equally satisfy the same description. In such cases, the task is inherently under-specified and is closely related to multi-instance grounding. Importantly, MVLM reduces this ambiguity through the vision-language margin confidence (κ_t^{mvlm}). When multiple sim-



(a) Language description: "*the target person the player against with*"

(b) Language description: "*the person who riding a bike on the road*"



(c) Language description: "*the cat in the cage*"

Figure 4. Qualitative comparison on three challenging sequences. MVLM (red box) more accurately follows the language-described target compared to SOTA trackers. SOTA trackers frequently drift to visually similar distractors or lose the target under occlusion, while MVLM preserves correct localization through margin-based correlation and confidence-driven ROI selection.

ilar targets exist, the correlation margin between the top-1 and top-2 candidates naturally becomes small. This correspondingly lowers the overall κ_t^{mvlm} score, which prevents over-confident drift by triggering a conservative search decision. Furthermore, as a straightforward extension, with minimal additional user disambiguation (e.g., a short temporal hint or an extra phrase), our framework can maintain top-K candidates and compare their temporal-memory margin confidences to progressively narrow down to the intended target.

5.2. More Qualitative Results

Figure 4 presents qualitative tracking results taken from diverse scenarios, including RGB game footage, urban thermal imagery, and crowded outdoor environments. Across these cases, MVLM consistently identifies the target described by natural language while suppressing visually similar distractors. In Figure 4-(a), the target undergoes significant pose changes and partial occlusions. MVLM maintains stable tracking by relying on correlation margin and memory-based confidence, preventing drift toward background characters. In Figure 4-(b), where the textual cue is weak and clutter levels are high, the tracker still localizes the described individuals. This demonstrates that vision-language correlation is robust across modalities. In Figure 4-(c), multiple people or animals appear with similar visual appearance. Even in such ambiguous layouts, MVLM discards low-confidence candidates and retains the correct one, showing the effectiveness of the ROI filtering guided by MVLM confidence.

5.3. Attribute-based Evaluation

To further analyze the behavior of MVLM under specific tracking conditions, we evaluate the tracker on 17 attribute subsets of TNL2K [5]. The OPE normalized precision (NPR) curves in Figure 5 compare our method against two

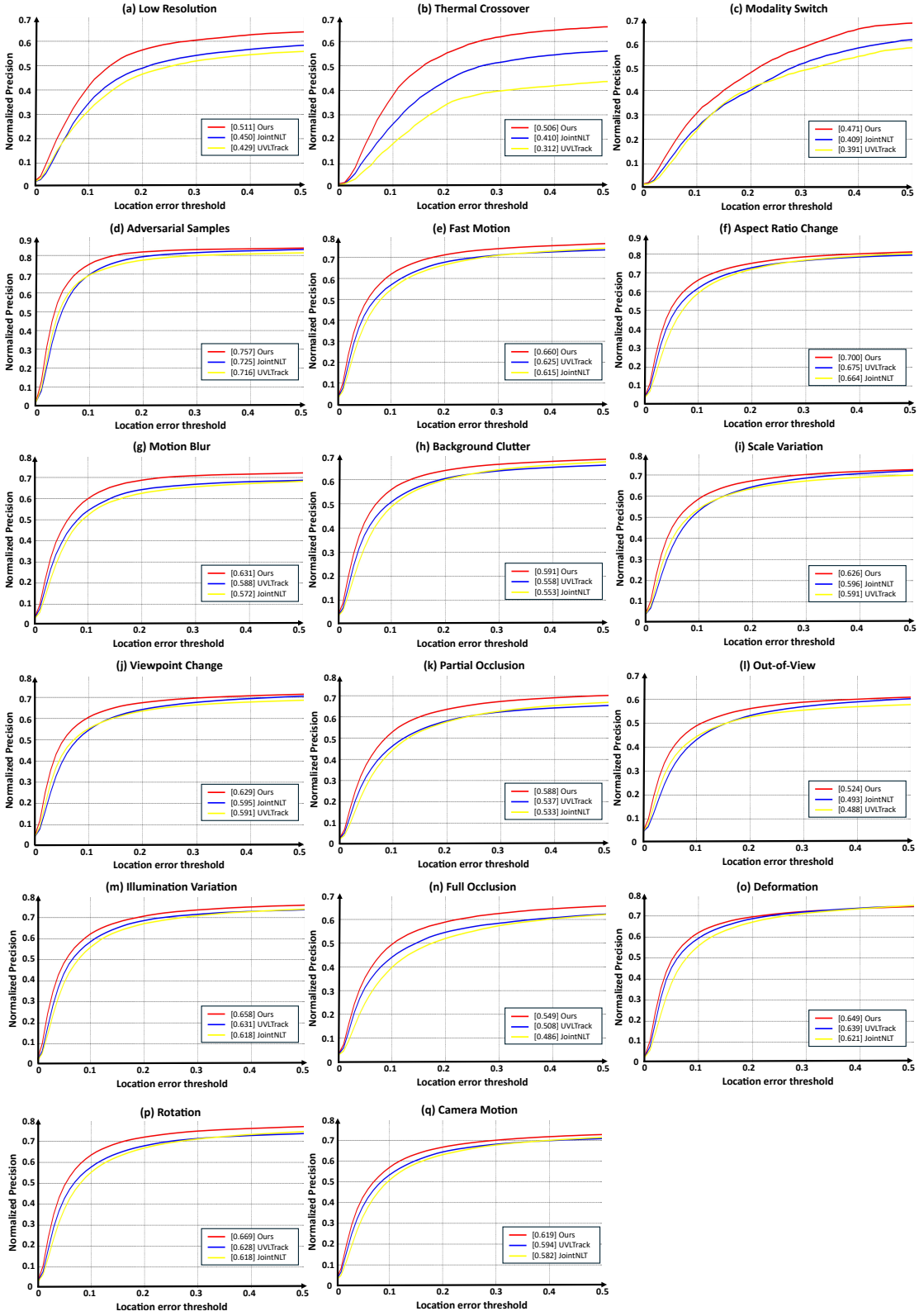


Figure 5. OPE normalized precision curves on the 17 attribute-based subsets of TNL2K [5]. Across these attributes, MVLM (red line) outperforms JointNLT [6] (blue line) and UVLTrack [3] (yellow line).

Method	FLOPs	Params	Speed
JointNLT [6]	42.2G	153M	39 FPS
UVLTrack [3]	33.1G	142M	53 FPS
MVLM (Ours)	24.7G	127M	56 FPS

Table 1. Comparison on computational cost.

SOTA vision-language trackers, JointNLT [6] and UVLTrack [3]. Across these attributes, MVLM consistently achieves the best NPR. On attributes on low resolution, thermal crossover and modality switch, MVLM outperforms NPR 6.1%, 9.6% and 6.2% compared to the second best tracker. This result demonstrates that even when the target appearance collapses or undergoes drastic variations, our model still preserves discriminability against distractors. This highlights that the correlation margin acts as a compensatory signal when visual cues degrade. We show that our margin signal preserves discriminability against distractors. On the other hand, NPR improvements are minor in attributes where the geometric variation is deformed but the target identity is still maintained. For example, attributes on aspect ratio change, scale variation, and deformation have the NPR gain 2.5%, 3.0%, 1.0% compared to the second-best tracker, respectively. These observations demonstrate that the correlation margin acts as a compensatory signal against appearance degradation. When the visual cue collapses, the margin becomes the dominant cue that preserves target identity. As a result, our proposed correlation-margin loss and MVLM effectively enhance discriminability under challenging attribute conditions.

6. Analysis for Computational Cost and Speed

In this section, we provide a detailed analysis of the computational cost and tracking speed of MVLM compared to state-of-the-art vision-language trackers. As summarized in Table 1, evaluated under the same hardware settings, our method requires 24.7G FLOPs and 127M parameters, running at a real-time speed of 56 FPS. In comparison, JointNLT requires 42.2G FLOPs and 153M parameters running at 39 FPS, while UVLTrack requires 33.1G FLOPs and 142M parameters running at 53 FPS. Beyond the network efficiency, our memory-gated tracking strategy further improves practical runtime efficiency. By dynamically switching to compact local ROI tracking whenever the MVLM confidence is high, the framework avoids redundant global feature extractions, invoking the heavier global search only when strictly necessary.

References

[1] 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. 5

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 4

[3] 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. 8, 9

[4] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019. 4

[5] 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. 7, 8

[6] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *CVPR*, pages 23151–23160, 2023. 8, 9