Supplementary Material: Triple-Reference Evaluation (TRE) for FGD-Det

1 Triple-Reference Evaluation

1.1 Motivation

Standard mAP in multi-spectral object detection evaluates predictions solely against the RGB-coordinate ground truth, making it unable to distinguish two fundamentally different model behaviors: (i) the model genuinely fuses both modalities; (ii) the model simply ignores IR and relies on RGB alone, thereby avoiding the negative impact of misalignment. When physical parallax exists between dual cameras in real-world deployment, the same mAP value can correspond to either scenario. TRE diagnoses fusion quality by actively applying controlled shifts to the IR modality and observing the model's behavioral response.

1.2 Three Reference Frames

For each annotated object with bounding box $\mathbf{b} = (x_1, y_1, x_2, y_2)$, three references are constructed at each offset magnitude $s = \sqrt{\Delta x^2 + \Delta y^2}$:

RGB-GT. The original ground-truth boxes, fixed regardless of IR offset:

$$\mathcal{G}_{\mathrm{RGB}} = \{\mathbf{b}_i\}_{i=1}^{N}$$

X-GT. The ground-truth boxes shifted by the same translation applied to the IR modality:

$$\mathcal{G}_{\mathrm{X}} = \left\{\mathbf{b}_i + \left(\frac{\Delta x}{W}, \frac{\Delta y}{H}, \frac{\Delta x}{W}, \frac{\Delta y}{H}\right)\right\}_{i=1}^{N}$$

where (W, H) is the image size.

Union-GT. The bounding box union of each RGB-GT and X-GT pair:

$$\mathcal{G}_{\cup} = \left\{\left(\min(x_1^i, x_1^{i\prime}),\ \min(y_1^i, y_1^{i\prime}),\ \max(x_2^i, x_2^{i\prime}),\ \max(y_2^i, y_2^{i\prime})\right)\right\}_{i=1}^{N}$$

where $\mathbf{b}_i' \in \mathcal{G}_{\mathrm{X}}$ is the shifted counterpart of $\mathbf{b}_i \in \mathcal{G}_{\mathrm{RGB}}$.

1.3 IR Shift Mechanism

To simulate physical sensor parallax during evaluation, IR channels are shifted via affine grid sampling. For each destination pixel coordinate $\mathbf{p}_{\text{dst}} = (u, v)$ in the output, the corresponding source coordinate is:

$$\mathbf{p}_{\text{src}} = \mathbf{p}_{\text{dst}} - \left(\frac{2\Delta x}{W}, \ \frac{2\Delta y}{H} \right)$$

The shifted IR feature map is obtained via differentiable bilinear interpolation:

$$I_{\text{shifted}}(u, v) = \sum_{m=0}^{1} \sum_{n=0}^{1} w_{mn} \cdot I\big(\lfloor u_s \rfloor + m, \ \lfloor v_s \rfloor + n\big)$$

where $(u_s, v_s) = \mathbf{p}_{\text{src}}$ and w_{mn} are the bilinear weights.

A "subtract gray → pad → add gray" strategy is employed for border regions, ensuring that out-of-bounds sampling regions are filled with the YOLO default gray value (114/255), matching the training-time `IRRandomShift` augmentation.

1.4 Behavior Classification

For each predicted box $\mathbf{p}$ with confidence ≥ 0.25, greedy IoU matching (IoU ≥ 0.5, class-consistent) is performed independently against $\mathcal{G}_{\text{RGB}}$, $\mathcal{G}_{\text{X}}$, and $\mathcal{G}_{\cup}$.

Each prediction is classified into one of four mutually exclusive categories, as summarized in Table 1:

Table 1 TRE behavior classification.

Behavior	**Condition**	**Interpretation**
RGB-Grounded	Matches $\mathcal{G}_{\text{RGB}}$	Safe; anchored to RGB
X-Grounded	Matches $\mathcal{G}_{\text{X}}$, not $\mathcal{G}_{\text{RGB}}$	Misled by IR offset
Fused	Matches $\mathcal{G}_{\cup}$, neither alone	Genuine cross-modal integration
Spurious	Matches none	IR-induced false positive

1.5 Metrics

Robust AP ($\mathbf{AP}_{\mathbf{robust}}$). Standard average precision against $\mathcal{G}_{\text{RGB}}$ under IR offset s, using 101-point interpolated AP:

$$\text{AP}_{\text{robust}}(s) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{101} \sum_{r \in \{0, 0.01, \ldots, 1.0\}} \max_{\tilde{r} \geq r} p_c(\tilde{r}, s)$$

where $\mathcal{C}$ is the set of object classes.

Modality Reliance Ratio (MRR).

$$\text{MRR}(s) = \frac{N_{\text{RGB-Grounded}}(s)}{N_{\text{X-Grounded}}(s)}$$

MRR $\gg 1$ indicates RGB-dominant localization with IR assistance (desired). When $N_{\text{X-Grounded}} = 0$, MRR is capped at 10^6 for numerical stability.

Spurious Detection Rate (SDR).

$$\text{SDR}(s) = \frac{N_{\text{Spurious}}(s)}{N_{\text{Total}}(s)}$$

where $N_{\text{Total}}(s)$ is the total number of predictions at confidence ≥ 0.25.

Robust Fusion Score (RFS).

$$\text{RFS}(s) = \text{AP}_{\text{robust}}(s) \cdot \frac{\ln\big(\text{MRR}(s) + 1\big)}{\text{SDR}(s) + \varepsilon}$$

where $\varepsilon = 10^{-7}$. The logarithmic scaling of MRR reflects diminishing returns beyond strong RGB dominance.

Integrated RFS ($\mathbf{RFS_{total}}$).

$$\text{RFS}_{\text{total}} = \int_0^{S_{\max}} \text{RFS}(s)\, ds \approx \sum_{k=1}^{K-1} \frac{\text{RFS}(s_k) + \text{RFS}(s_{k+1})}{2} \cdot (s_{k+1} - s_k)$$

where RFS(s) is sampled at $s \in \{3, 6, 9, 12, 15\}$ pixels. $s = 0$ is excluded because behavior classification is degenerate at zero offset.

1.6 Implementation

The TRE evaluation pipeline is implemented as `TREValidator` within the Ultralytics framework. For each offset, IR channels are shifted via affine grid sampling, model inference is run, predictions are post-processed, and per-image behavior classification is performed using NumPy matching routines. A full run over 6 offsets on M3FD ($\sim$6,200 images) completes in approximately 15 minutes on a single GPU. Figure 1 provides a conceptual illustration of the TRE framework.

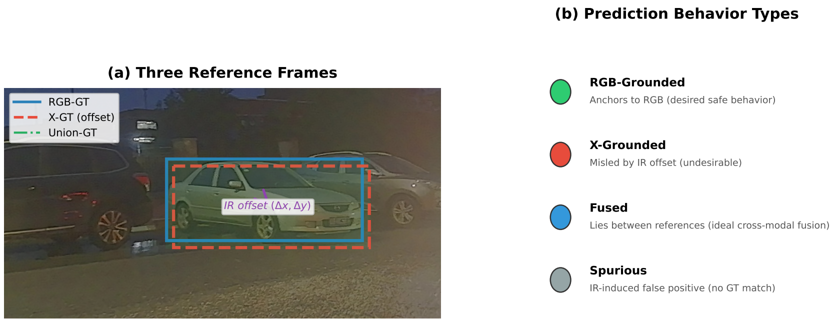

Fig. 1 Illustration of the TRE framework. (a) Three reference frames (RGB-GT, X-GT, Union-GT) for the same target under IR offset. (b) Four prediction behavior types with color coding.