TARGET-CENTRIC FREQUENCY DISTILLATION: LEARNING SPECTRALLY ALIGNED REPRESENTATIONS FOR RGB-T VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB-Thermal (RGB-T) visual grounding localizes a text-referred object in paired RGB and infrared (IR) images. However, RGB-centric vision-language models provide limited IR semantic modeling, hindering the integration of IR cues into query-conditioned representations. To address this, we propose Target-Centric Frequency Distillation (TCFDistill), a framework for learning spectrally aligned representations for robust grounding. Our key insight is that query-relevant RGB–IR structures exhibit distinct frequency characteristics: low frequencies capture slowly varying, coarse-scale spatial structures, while high frequencies reflect fine-grained local variations. Specifically, we introduce Dual-Frequency Spectral Distillation (DFSD), which refines IR representations through structural knowledge transfer from a pretrained RGB–IR teacher model. DFSD aligns query-conditioned low-frequency target-affinity relations and high-frequency RGB–IR correspondences at matched positions and within local neighborhoods. Furthermore, Spectral Alignment-Guided Representation Modulation (SARM) uses local cross-modal frequency consistency to modulate shared structures and complementary responses between RGB and refined IR representations, producing spatially organized frequency tokens for grounding. Experiments on RefFLIR, RefM^{3}FD, and RefMFAD demonstrate state-of-the-art performance, surpassing RGBT-VGNet by 2.79% on RefM^{3}FD Test-B and 4.54% on RefMFAD Test-C. All code will be released upon acceptance.

1 INTRODUCTION

Visual grounding (VG) localizes the object referred to by a textual expression, supporting applications such as language-guided navigation and human–robot interaction Deng et al. (2021); Xiao et al. (2026). Existing methods primarily rely on RGB images, whose color, texture, and appearance cues can degrade under low illumination, strong glare and occlusion imaging Hnewa et al. (2023). RGB-Thermal (RGB-T) visual grounding incorporates infrared (IR) images, which are less sensitive to illumination changes, to complement RGB observations with thermal contrast, object contours, and local structures for more reliable localization in challenging environments Zhao et al. (2026).

Visual grounding methods commonly employ pretrained vision-language models (VLMs), such as CLIP Radford et al. (2021), as representation encoders. However, these models are pretrained primarily on RGB image–text pairs and provide limited support for IR semantic modeling, constraining the integration of IR cues into query-conditioned representations. Existing RGB-T visual grounding models integrate RGB and IR cues at the pixel, feature, or decision level, while RGBT-VGNet further incorporates language cues and modality reliability into the grounding process Yuan et al. (2025); Zhao et al. (2026). Although these models facilitate the use of complementary IR cues, they do not explicitly distill query-relevant RGB–IR structural relations.

Cross-modal distillation offers a way to enhance IR representations with teacher knowledge. Existing methods typically transfer semantic representations or localization predictions from an RGB teacher model to an IR student model Ranasinghe et al. (2026), without explicitly modeling RGB–IR structural relations. As shown in Fig. 1(a), a pretrained RGB–IR teacher model (e.g., UNIV Mao et al. (2025)) achieves a higher structural consistency score than the VLM, indicating better preserva-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

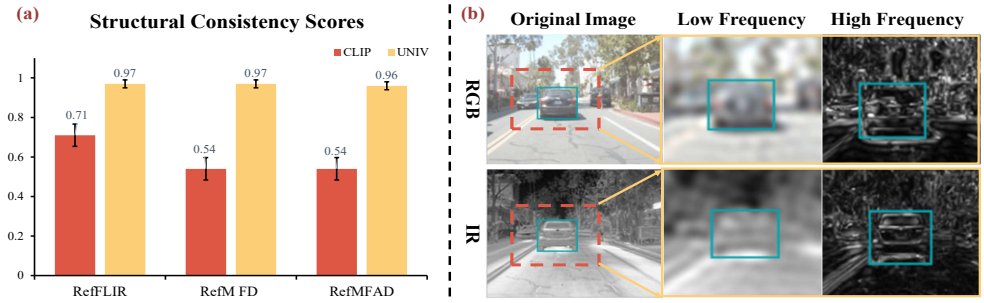


Figure 1: Motivation for learning spectrally aligned representations in RGB-T visual grounding. (a) Comparison of cross-modal structural consistency[1] between a representative VLM (e.g., CLIP) and an RGB–IR pretrained teacher model (e.g., UNIV). Higher scores indicate better preservation of RGB spatial similarity patterns in cross-modal RGB–IR relations. (b) Frequency decomposition of representative RGB–IR pairs. Low-frequency components capture slowly varying, coarse-scale spatial structures, whereas high-frequency components reflect fine-grained local variations.

tion of RGB spatial similarity patterns in cross-modal RGB–IR relations. Transferring such relations facilitates IR refinement and the integration of complementary IR cues into query-conditioned representations. However, the teacher model and the VLM differ in pretraining objectives and representation spaces, making direct feature matching potentially suboptimal and motivating relation-level transfer. Such structural relations can be further organized across spatial scales through frequency decomposition. As illustrated in Fig. 1 (b), low-frequency components capture slowly varying, coarse-scale spatial structures, whereas high-frequency components reflect fine-grained local variations. However, existing frequency-aware distillation methods, such as FreeKD Zhang et al. (2024) and SDKD Li et al. (2025), exploit frequency information without identifying which RGB–IR structural relations are relevant to a given text query, motivating target-centric structural distillation across frequency bands.

These observations reveal two key challenges for RGB-T visual grounding: transferring frequency-specific structural knowledge across heterogeneous representation spaces, and aligning shared RGB–IR structures while retaining modality-specific complementary information. To address these challenges, we propose Target-Centric Frequency Distillation (TCFDistill), a framework for learning spectrally aligned representations through Dual-Frequency Spectral Distillation (DFSD) and Spectral Alignment-Guided Representation Modulation (SARM). DFSD distills target-centric structural knowledge from the teacher model into the student branch through two complementary objectives: Low-Frequency Target-Affinity Distillation (LF-TAD) and High-Frequency Structural Correspondence Distillation (HF-SCD). LF-TAD uses RGB–text relevance weights to construct query-conditioned target prototypes from low-frequency IR features and aligns the resulting target-affinity relations. HF-SCD aligns high-frequency RGB–IR correspondences at matched positions and within local neighborhoods. Both objectives transfer structural relations without directly matching feature values between the teacher model and the student branch. SARM then uses local cross-modal frequency consistency to modulate shared structures and complementary responses between RGB and refined IR representations, producing spatially organized frequency tokens for grounding.

Experiments on RefFLIR, RefM^{3}FD, and RefMFAD demonstrate the effectiveness of the proposed framework across diverse imaging conditions, with particularly notable improvements under low illumination and for small targets. Our contributions are summarized as follows:

- We propose TCFDistill, a framework for learning spectrally aligned representations through frequency-specific structural alignment and modulation.
- We introduce DFSD to align query-conditioned low-frequency target affinities and high-frequency local RGB–IR correspondences, and SARM to generate spatially organized frequency tokens through consistency-guided modulation.
- Experiments on RefFLIR, RefM^{3}FD, and RefMFAD demonstrate the effectiveness of TCFDistill, with notable gains under low illumination and for small targets.

[1]The computation of the structural consistency score is detailed in Appendix B.

2 Related Work

2.1 Visual Grounding and RGB-T Grounding

Recent progress in visual and language representation learning has substantially advanced visual grounding Deng et al. (2021); Ye et al. (2022); Xiao et al. (2023; 2024a;b); Wang et al. (2025); Kang et al. (2025); Wang et al. (2026); Hu et al. (2025). Transformer-based methods, such as TransVG Deng et al. (2021) and QRNet Ye et al. (2022), improve visual-language interaction through end-to-end regression and query-aware visual refinement, respectively. Building on pre-trained vision-language models, CLIP-VG Xiao et al. (2023), HiVG Xiao et al. (2024a), and OneRef Xiao et al. (2024b) further enhance grounding through CLIP adaptation, hierarchical multimodal modulation, and unified multimodal representation learning. Most of these methods operate on RGB images. RGBT-GroundBench Zhao et al. (2026) extends grounding to RGB–IR image pairs and introduces RGBT-VGNet to fuse the two modalities according to language cues and modality reliability. TCFDistill instead combines teacher-guided transfer of query-relevant structures across frequency bands with local modulation of shared structures and complementary responses based on cross-modal frequency consistency.

2.2 Frequency-Aware Cross-Modal Knowledge Distillation

Knowledge distillation has evolved from direct feature imitation Zagoruyko & Komodakis (2017) toward structured relation transfer, self-distillation Wu et al. (2024), frequency-aware supervision, and cross-modal knowledge transfer Huang et al. (2022). Frequency-domain distillation decomposes teacher representations into spectral components and introduces frequency-specific transfer objectives Zhang et al. (2024); Li et al. (2025). FreeKD Zhang et al. (2024) uses semantic frequency prompts and position-aware relation matching to select informative frequency responses, while Frequency-Aligned Knowledge Distillation Li et al. (2025) separately transfers low-frequency global trends and high-frequency local variations. Cross-modal distillation transfers complementary knowledge between heterogeneous modalities and has been explored in RGB, depth, and infrared perception Bao et al. (2026); Lu et al. (2024); Liu et al. (2021); Liu & Zhang (2025). Recent work also explores frequency-disentangled feature alignment for cross-modal knowledge transfer Liu et al. (2026). Different from feature-level frequency alignment, TCFDistill transfers target-centric spatial relations: LF-TAD distills query-conditioned low-frequency target affinities, whereas HF-SCD distills local high-frequency RGB–IR correspondences without directly regressing heterogeneous teacher and student features.

3 Methodology

3.1 Overview

Given an RGB–IR image pair $(\boldsymbol{I}^{rgb}, \boldsymbol{I}^{ir})$ and a referring expression Q, we propose **Target-Centric Frequency Distillation (TCFDistill)** to learn spectrally aligned representations. As illustrated in Fig. 2, it comprises two complementary components: **Dual-Frequency Spectral Distillation (DFSD)** transfers target-centric, frequency-specific structural relations from the frozen teacher model to the student branch for IR refinement, while **Spectral Alignment-Guided Representation Modulation (SARM)** uses local cross-modal frequency consistency to modulate shared structures and complementary responses, producing spatially organized frequency tokens.

We first encode the images and referring expression using the VLM's visual encoder $\mathrm{Encoder}_{\mathrm{vis}}$ and text encoder $\mathrm{Encoder}_{\mathrm{text}}$:

$$\boldsymbol{Z}^{rgb} = \mathrm{Encoder}_{\mathrm{vis}}\left(\boldsymbol{I}^{rgb}\right), \qquad \boldsymbol{Z}^{ir} = \mathrm{Encoder}_{\mathrm{vis}}\left(\boldsymbol{I}^{ir}\right), \qquad \boldsymbol{Z}^{q} = \mathrm{Encoder}_{\mathrm{text}}(Q), \tag{1}$$

where $\boldsymbol{Z}^{rgb}$ and $\boldsymbol{Z}^{ir}$ denote the visual representations, which are aggregated and projected for grounding. $\boldsymbol{Z}^{q}$ denotes the text representations used for query guidance and grounding.

During training, DFSD uses representations $\boldsymbol{T}^{rgb}$ and $\boldsymbol{T}^{ir}$ from the teacher model to supervise the student branch. This supervision combines query-conditioned low-frequency target-affinity alignment with high-frequency RGB–IR correspondence alignment. The student branch applies a residual adapter to the gradient-detached IR representation, producing $\boldsymbol{Z}^{aux}$. An RGB–text relevance

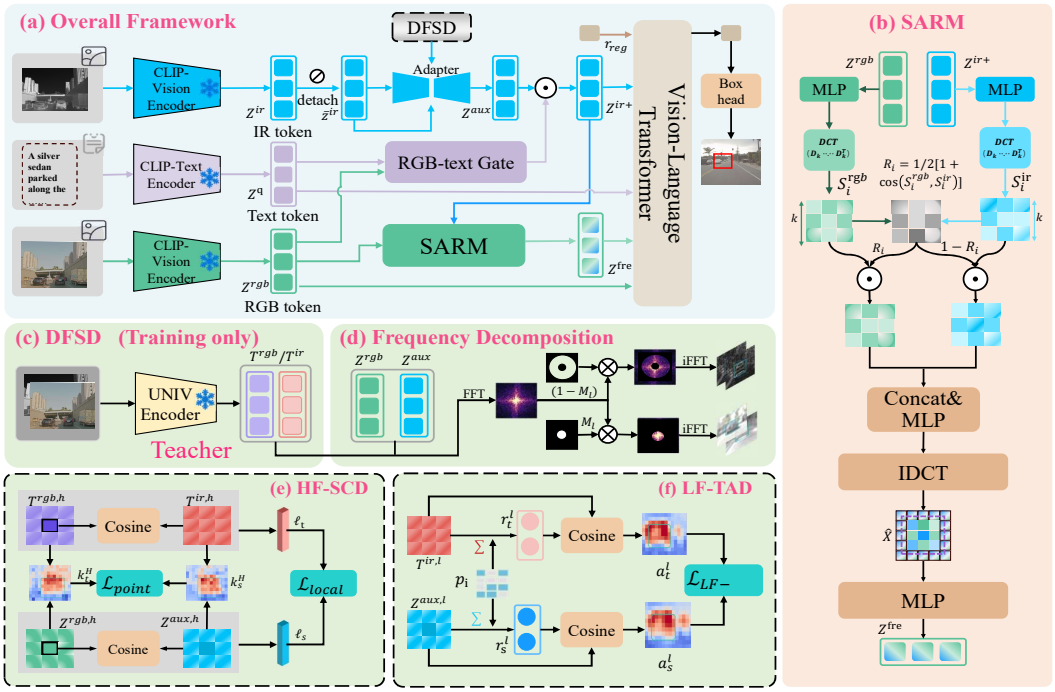


Figure 2: Overall framework of Target-Centric Frequency Distillation (TCFDistill), comprising Dual-Frequency Spectral Distillation(DFSD) and Spectral Alignment-Guided Representation Modulation (SARM). DFSD consists of Low-Frequency Target Affinity Distillation (LF-TAD) and High-Frequency Structural Correspondence Distillation (HF-SCD).

gate selectively injects the projected auxiliary representation into the IR stream:

$$\boldsymbol{Z}^{ir+} = \boldsymbol{Z}^{ir} + G\left(\boldsymbol{Z}^{rgb}, \boldsymbol{Z}^{q}\right) \odot \phi_0\left(\boldsymbol{Z}^{aux}\right), \tag{2}$$

where $G(\cdot)$ is a parameter-free RGB–text relevance gate, and $\phi_0(\cdot)$ is a zero-initialized projection. Its computation is given in Appendix D.5.

SARM performs local spectral interaction and consistency-conditioned modulation between the RGB and refined IR representations, producing spatially organized frequency tokens:

$$\boldsymbol{Z}^{fre} = \mathrm{SARM}\left(\boldsymbol{Z}^{rgb}, \boldsymbol{Z}^{ir+}\right). \tag{3}$$

The RGB, refined IR, frequency, and text tokens are then projected into a shared hidden space and concatenated with a learnable regression token ($\boldsymbol{r}_{\mathrm{reg}}$):

$$\boldsymbol{X}_0 = \left[\boldsymbol{r}_{\mathrm{reg}}; P_{\mathrm{rgb}}\left(\boldsymbol{Z}^{rgb}\right); P_{\mathrm{ir}}\left(\boldsymbol{Z}^{ir+}\right); P_{\mathrm{fre}}\left(\boldsymbol{Z}^{fre}\right); P_{\mathrm{q}}(\boldsymbol{Z}^{q})\right]. \tag{4}$$

A vision–language Transformer jointly processes these tokens and predicts the target bounding box from the updated regression token. The teacher is removed at inference.

3.2 Dual-Frequency Spectral Distillation

Dual-Frequency Spectral Distillation (DFSD) transfers target-centric, frequency-specific structural relations from the teacher model to the student branch, without directly matching teacher and student feature values. Let $\boldsymbol{Z}^{rgb}, \boldsymbol{Z}^{ir} \in \mathbb{R}^{N\times d}$ denote the RGB and IR representations produced by the pretrained VLM visual encoder. The teacher model extracts the corresponding RGB and IR representations, which are projected to obtain $\boldsymbol{T}^{rgb}, \boldsymbol{T}^{ir} \in \mathbb{R}^{N\times d}$.

IR Refinement Adapter. The student branch incorporates a IR refinement adapter. To prevent DFSD from directly updating the pretrained IR features, we detach the input representation: $\bar{\boldsymbol{Z}}^{ir} = \mathrm{sg}(\boldsymbol{Z}^{ir})$, where $\mathrm{sg}(\cdot)$ denotes the detach operation. The auxiliary representation is constructed as

$$\Delta \boldsymbol{Z}^{ir} = \boldsymbol{f}_{\mathbf{adapter}}\left(\bar{\boldsymbol{Z}}^{ir}\right), \qquad \boldsymbol{Z}^{aux} = \bar{\boldsymbol{Z}}^{ir} + \Delta \boldsymbol{Z}^{ir}, \tag{5}$$

where $\boldsymbol{f}_{\mathbf{adapter}}$ comprises a low-rank down-projection and an up-projection. DFSD supervises $\boldsymbol{Z}^{aux}$ through frequency-specific structural objectives. The auxiliary output is gated and injected into the IR stream through Eq. 2, yielding the refined IR representation used by SARM.

Frequency Decomposition. We reshape patch features into spatial grids, apply complementary radial masks in the Fourier domain to extract low- and high-frequency components, preserving the patch-grid layout. This yields $(\boldsymbol{Z}^{aux,l}, \boldsymbol{T}^{ir,l})$ for LF-TAD and $(\boldsymbol{Z}^{rgb,h}, \boldsymbol{Z}^{aux,h}, \boldsymbol{T}^{rgb,h}, \boldsymbol{T}^{ir,h})$ for HF-SCD. Implementation details are provided in Appendix A.3.

Low-Frequency Target-Affinity Distillation. Let $\boldsymbol{z}_i^{rgb} \in \mathbb{R}^d$ denote the RGB feature at position $i \in \{1, \ldots, N\}$. For simplicity, padding masks are omitted from the notation. We obtain a global query representation $\bar{\boldsymbol{z}}^q = \mathrm{Pool}(\boldsymbol{Z}^q)$, where $\mathrm{Pool}(\cdot)$ denotes CLIP's EOT pooling followed by the text projection. We then compute a query-conditioned relevance distribution over the RGB positions:

$$p_i = \frac{\exp\left(\cos\left(\boldsymbol{z}_i^{rgb}, \bar{\boldsymbol{z}}^q\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\cos\left(\boldsymbol{z}_j^{rgb}, \bar{\boldsymbol{z}}^q\right)/\tau\right)}, \tag{6}$$

where τ is the temperature parameter. Since RGB representations retain reliable language alignment, the resulting distribution $\boldsymbol{p} = [p_i]_{i=1}^N$ provides soft spatial weighting of query-relevant regions. Using the same response distribution to maintain a common query-conditioned spatial focus, we construct low-frequency target prototypes for the student and teacher:

$$\boldsymbol{r}_s^l = \sum_{i=1}^{N} p_i \boldsymbol{z}_i^{aux,l}, \qquad \boldsymbol{r}_t^l = \sum_{i=1}^{N} p_i \boldsymbol{t}_i^{ir,l}, \tag{7}$$

where $\boldsymbol{z}_i^{aux,l}$ and $\boldsymbol{t}_i^{ir,l}$ denote the low-frequency features of the IR refinement adapter and teacher, respectively. The student and teacher target affinities are then defined as

$$a_s^l(i) = \cos\left(\boldsymbol{z}_i^{aux,l}, \boldsymbol{r}_s^l\right), \qquad a_t^l(i) = \cos\left(\boldsymbol{t}_i^{ir,l}, \boldsymbol{r}_t^l\right), \tag{8}$$

where target affinity measures the similarity of each low-frequency feature to its corresponding query-conditioned target prototype. The LF-TAD objective aligns the resulting target-affinity maps:

$$\mathcal{L}_{\text{LF-TAD}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{SmoothL1}\left(a_s^l(i), a_t^l(i)\right). \tag{9}$$

LF-TAD matches affinity maps to transfer query-conditioned spatial structure across heterogeneous teacher and student feature spaces.

High-Frequency Structural Correspondence Distillation. We first measure same-position RGB–IR correspondence in the high-frequency student and teacher representations:

$$\kappa_s(i) = \cos\left(\boldsymbol{z}_i^{rgb,h}, \boldsymbol{z}_i^{aux,h}\right), \qquad \kappa_t(i) = \cos\left(\boldsymbol{t}_i^{rgb,h}, \boldsymbol{t}_i^{ir,h}\right), \tag{10}$$

where $\kappa_s(i)$ and $\kappa_t(i)$ denote the same-position correspondence scores of the student and teacher, respectively. Here, structural correspondence is quantified by feature similarity. Same-position correspondence alone provides limited information about the surrounding high-frequency structure. We therefore introduce a local correspondence term. Let $\mathcal{N}(i)$ denote the positions within the 3×3 neighborhood centered at i. The local correspondence of each RGB center feature is computed by averaging its similarities to the IR features within $\mathcal{N}(i)$:

$$\ell_s(i) = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \cos\left(\boldsymbol{z}_i^{rgb,h}, \boldsymbol{z}_j^{aux,h}\right), \qquad \ell_t(i) = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \cos\left(\boldsymbol{t}_i^{rgb,h}, \boldsymbol{t}_j^{ir,h}\right). \tag{11}$$

Compared with a dense pairwise correspondence matrix, this formulation introduces one local descriptor for each center position while retaining local structural information. Let Ω denote the spatial positions within the ground-truth target. The same-position and local correspondence objectives are defined as:

$$\mathcal{L}_{\text{point}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \mathrm{SmoothL1}\left(\kappa_{\mathrm{s}}(i), \kappa_{\mathrm{t}}(i)\right), \qquad \mathcal{L}_{\text{local}} = \frac{1}{|\Omega|} \sum_{i \in \Omega} \mathrm{SmoothL1}\left(\ell_{\mathrm{s}}(i), \ell_{\mathrm{t}}(i)\right). \tag{12}$$

The HF-SCD objective is formulated as:

$$\mathcal{L}_{\text{HF-SCD}} = \mathcal{L}_{\text{point}} + \lambda_{\text{local}}\mathcal{L}_{\text{local}}, \tag{13}$$

where λ_{local} balances the local correspondence term. HF-SCD uses teacher representations as fixed references and optimizes only the IR refinement adapter. Together, LF-TAD and HF-SCD transfer frequency-specific structural relations without directly matching teacher and student feature values.

3.3 Spectral Alignment-Guided Representation Modulation

As illustrated in Fig. 2, Spectral Alignment-Guided Representation Modulation (SARM) performs local spectral interaction and consistency-guided modulation between the RGB and DFSD-refined IR representations, producing spatially organized frequency tokens for grounding. Specifically, it decomposes spatially corresponding local neighborhoods in the RGB and refined IR representations into frequency components and uses their cross-modal frequency consistency to modulate shared structures and complementary responses.

Shared Projection and Local Spectral Decomposition. We project the RGB features $\boldsymbol{Z}^{rgb}$ and refined IR features $\boldsymbol{Z}^{ir+}$ through a shared MLP, obtaining $\boldsymbol{X}^{rgb}, \boldsymbol{X}^{ir} \in \mathbb{R}^{N\times d_f}$. The projected features are rearranged into $H \times W$ spatial grids, where $N = H \times W$. For each spatial position i, we extract paired $K \times K$ neighborhoods from the RGB and IR grids, denoted by $\mathcal{W}_i(\boldsymbol{X}^{rgb})$ and $\mathcal{W}_i(\boldsymbol{X}^{ir})$, respectively. We set $K = 3$ by default. Unlike the global FFT used in DFSD for frequency-band masking, SARM adopts the Discrete Cosine Transform (DCT) within local neighborhoods. Its real-valued, orthogonal basis enables direct cross-modal frequency modulation and subsequent reconstruction in the spatial domain. We apply a channel-wise two-dimensional DCT within each neighborhood:

$$\boldsymbol{S}_i^m = \boldsymbol{D}_K\, \mathcal{W}_i(\boldsymbol{X}^m)\, \boldsymbol{D}_K^\top, \qquad m \in \{rgb, ir\}, \tag{14}$$

where $\boldsymbol{D}_K \in \mathbb{R}^{K\times K}$ is the orthonormal DCT basis. The resulting tensor $\boldsymbol{S}_i^m \in \mathbb{R}^{K\times K\times d_f}$ contains K^2 local frequency coefficients. Each coefficient set remains associated with its center position i, thereby preserving the spatial organization of the patch features.

Consistency-Conditioned Spectral Modulation. Given the local RGB and IR frequency coefficients $\boldsymbol{S}_i^{rgb}$ and $\boldsymbol{S}_i^{ir}$, SARM first estimates their cross-modal consistency at each frequency:

$$\boldsymbol{R}_i = \frac{1}{2}\left[1 + \cos_c\left(\boldsymbol{S}_i^{rgb}, \boldsymbol{S}_i^{ir}\right)\right], \qquad \boldsymbol{R}_i \in [0,1]^{K\times K}, \tag{15}$$

where $\cos_c$ denotes channel-wise cosine similarity. Each entry $\boldsymbol{R_i}$ is the rescaled RGB–IR cosine similarity at local frequency with larger values indicating greater cross-modal consistency. The consistency map then modulates the two spectral streams:

$$\boldsymbol{H}_i = \text{Concat}_c\left[\boldsymbol{R}_i \odot \boldsymbol{S}_i^{rgb}, (1-\boldsymbol{R}_i) \odot \boldsymbol{S}_i^{ir}\right], \tag{16}$$

where Concat_c denotes channel-wise concatenation and $\odot$ denotes frequency-wise multiplication. The consistency map is broadcast along the channel dimension. The first term emphasizes RGB components consistent with IR, whereas the second highlights IR components under greater cross-modal discrepancy. A lightweight modulation function subsequently integrates these components:

$$\widetilde{\boldsymbol{H}}_i = \psi\left(\boldsymbol{H}_i\right), \tag{17}$$

where $\psi(\cdot)$ is a two-layer MLP with GELU that projects $2d_f$-dimensional inputs to d_f dimensions. SARM retains all K^2 DCT coefficients and applies learned modulation at each frequency rather than manually selecting frequency bands.

Local Reconstruction and Representation Projection. The modulated frequency coefficients are transformed back into the local spatial domain using the inverse DCT:

$$\widehat{\boldsymbol{X}}_i = \boldsymbol{D}_K^\top \widetilde{\boldsymbol{H}}_i \boldsymbol{D}_K. \tag{18}$$

The reconstructed neighborhoods are restored to their original grid positions, and overlapping contributions are averaged to obtain $\widehat{\boldsymbol{X}} \in \mathbb{R}^{N\times d_f}$. An MLP projection then produces the SARM representation $\boldsymbol{Z}^{fre} \in \mathbb{R}^{N\times d}$, consisting of N spatially organized frequency tokens. This representation is concatenated with the RGB, refined IR, and query representations for grounding.

Table 1: Comparison with representative visual grounding methods on RefFLIR, RefM^{3}FD, and RefMFAD under single-modality and multimodal RGB-T settings. The best and second-best RGB–T results are shown in bold and underlined, respectively; "–" indicates unavailable results.

Method	Venue	Input Modality	RefFLIR					RefM^{3}FD					RefMFAD				
			Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
(a) Single-Modality Visual Grounding Methods (RGB or IR)																	
D-MDETR Shi et al. (2023)	TPAMI'24	RGB	57.07	49.25	70.17	39.22	26.77	52.38	50.19	71.67	57.04	26.08	60.59	59.39	86.71	55.79	40.98
MMCA Yao et al. (2024)	ACMMM'24		50.16	46.57	71.48	35.63	21.50	-	-	-	-	-	47.85	46.52	76.71	43.84	26.24
FSVG Wang et al. (2025)	ICME'25		56.09	45.25	67.14	36.88	20.55	49.40	49.39	73.86	54.65	23.70	59.71	58.65	86.24	55.34	39.81
AttBalance Kang et al. (2025)	ACMMM'25		51.64	49.10	73.81	40.16	21.76	-	-	-	-	-	51.04	51.30	83.92	48.74	29.61
HiVG-B Xiao et al. (2024a)	ACMMM'24		69.08	66.65	88.81	54.06	43.52	69.24	68.15	89.40	56.09	45.04	65.45	64.02	90.66	60.52	45.57
OneRef-B Xiao et al. (2024b)	NeurIPS'24		63.82	61.69	84.73	52.34	35.90	66.07	66.66	94.64	70.93	39.07	62.34	60.61	90.76	57.63	40.86
D-MDETR Shi et al. (2023)	TPAMI'24	IR	48.36	38.87	55.83	37.66	18.38	47.62	50.34	70.13	57.68	28.52	56.37	54.52	79.72	51.63	37.73
MMCA Yao et al. (2024)	ACMMM'24		48.52	41.41	60.98	40.00	18.46	-	-	-	-	-	46.26	45.22	76.33	42.86	24.43
FSVG Wang et al. (2025)	ICME'25		47.86	38.50	54.64	37.81	17.13	52.98	51.48	72.24	60.31	27.00	54.54	54.52	79.92	52.81	36.05
AttBalance Kang et al. (2025)	ACMMM'25		43.75	43.39	66.83	39.22	15.92	33.93	37.63	64.44	45.80	8.77	45.30	45.45	78.61	43.35	22.43
HiVG-B Xiao et al. (2024a)	ACMMM'24		68.75	64.07	81.55	66.09	42.41	68.09	61.77	78.21	68.31	40.59	63.22	62.63	87.50	60.85	45.69
OneRef-B Xiao et al. (2024b)	NeurIPS'24		62.01	60.00	80.55	60.00	35.80	60.12	65.10	87.74	73.03	41.02	58.60	57.95	86.46	55.71	38.67
(b) Multimodal RGB-T Visual Grounding Methods (RGB + IR)																	
MV-D-MDETR Shi et al. (2023)	TPAMI'24	RGB+IR	55.26	47.42	68.33	41.25	23.38	48.81	46.50	65.91	53.82	24.19	58.76	57.31	83.84	53.79	39.30
MV-MMCA Yao et al. (2024)	ACMMM'24		54.93	48.97	72.43	42.34	22.62	46.43	47.83	72.16	56.76	20.45	53.62	54.41	84.41	51.47	34.16
MV-FSVG Wang et al. (2025)	ICME'25		54.61	47.96	68.57	40.16	24.29	47.62	46.32	64.61	54.74	25.00	58.44	58.60	85.86	55.18	40.05
MV-AttBalance Kang et al. (2025)	ACMMM'25		55.26	54.04	76.07	48.59	27.63	46.43	50.50	79.30	58.39	20.40	52.71	52.69	82.58	50.94	31.74
MV-HiVG-B Xiao et al. (2024a)	ACMMM'24		75.33	69.20	85.80	66.72	50.20	69.64	72.35	93.99	79.43	49.40	67.07	67.04	91.01	64.52	51.53
MV-OneRef-B Xiao et al. (2024b)	NeurIPS'24		63.82	62.05	86.19	57.34	35.43	62.50	63.98	89.20	69.34	37.93	61.62	59.97	89.25	56.85	41.58
RGBT-VGNet Zhao et al. (2026)	ECCV'2026		75.33	72.43	89.64	69.22	52.53	73.21	74.89	94.16	80.75	55.52	68.31	67.58	91.29	64.80	51.18
TCFDistill	Ours	RGB+IR	**77.13**	**75.00**	**91.88**	**69.84**	**55.57**	**77.38**	**77.33**	**95.45**	**83.54**	**58.87**	**72.05**	**70.40**	**91.64**	**67.33**	**55.72**

Table 2: Comparison of distillation targets on RefM^{3}FD.

Variant	RefM^{3}FD				
	Val	Test	Test-A	Test-B	Test-C
Direct feature	76.19	75.52	**95.61**	81.71	55.41
Frequency feature	76.19	74.94	95.21	81.07	54.76
Relation (DFSD)	**77.38**	**77.33**	95.45	**83.54**	**58.87**

Table 3: Refinement modality ablation on RefM^{3}FD.

Refined modality	RefM^{3}FD				
	Val	Test	Test-A	Test-B	Test-C
RGB only	76.19	76.90	95.37	82.81	57.44
RGB + IR	76.78	76.77	**95.53**	82.54	58.00
IR only (Full)	**77.38**	**77.33**	95.45	**83.54**	**58.87**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets and Evaluation Metrics. Following the experimental protocol of RGBT-GroundBench Zhao et al. (2026); Xiao et al. (2026), we evaluate our method on three RGB-T visual grounding datasets Zhao et al. (2026): RefFLIR, RefM^{3}FD, and RefMFAD. These datasets contain paired RGB and IR images, referring expressions, and object bounding-box annotations under diverse environmental conditions. We adopt the official training and evaluation splits, including the validation set and the test, Test-A, Test-B, and Test-C subsets. Test-B covers weak-light nighttime scenes, while Test-C focuses on small targets. We evaluate Acc@0.5, where a prediction is correct if its IoU is at least 0.5. Further details are provided in Appendix A.2.

Implementation Details. Our model is built on HiVG Xiao et al. (2024a) with CLIP Radford et al. (2021) ViT-B/16 d'Ascoli et al. (2021) as the visual-language backbone. The RGB and IR images are resized to 224×224. We adopt the base-sized UNIV encoder as teacher model, with a ConvViT-B/16 backbone d'Ascoli et al. (2021) at the same resolution. We train the model for 120 epochs using AdamW with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. The learning rate is adjusted using a cosine schedule. Random cropping, scaling, and translation are applied for data augmentation. All experiments are conducted on a server equipped with two NVIDIA A100 GPUs, using a batch size of 32 per GPU. More details are provided in the Appendix A.

Table 4: Component ablation on RefM^{3}FD.

Variant	SARM	DFSD	RefM^{3}FD				
			Val	Test	Test-A	Test-B	Test-C
A: Baseline			72.02	75.13	94.23	80.98	55.62
B: + SARM	✓		75.00	76.66	95.26	82.35	57.19
C: + DFSD		✓	74.40	76.11	95.37	82.54	56.54
D: TCFDistill (Full)	✓	✓	**77.38**	**77.33**	**95.45**	**83.54**	**58.87**

Table 5: DFSD ablation on RefM^{3}FD.

Variant	LF	HF	Loss	RefM^{3}FD				
				Val	Test	Test-A	Test-B	Test-C
LF only	✓		SmoothL1	76.19	76.11	95.21	82.54	56.65
HF only		✓	SmoothL1	**77.97**	76.85	95.12	82.17	58.49
LF + HF	✓	✓	MSE	75.59	77.11	**95.86**	82.44	58.54
LF + HF	✓	✓	SmoothL1	77.38	**77.33**	95.45	**83.54**	**58.87**

Table 6: Frequency assignment in DFSD on RefM^{3}FD.

Variant	RefM^{3}FD				
	Val	Test	Test-A	Test-B	Test-C
HF-TAD only	75.00	76.11	**95.61**	83.27	56.38
LF-SCD only	76.78	76.00	95.45	81.99	56.54
Full-spectrum	74.40	76.32	95.21	82.63	57.03
Swapped	76.78	77.03	95.04	82.54	57.03
TCFDistill (Full)	**77.38**	**77.33**	95.45	**83.54**	**58.87**



Figure 3: Comparison of query-conditioned heatmaps in challenging RGB–IR scenes.

4.2 Comparison with State-of-the-Art Methods

Comparison with Single- and Multimodal Methods. Table 1 compares representative methods under single- and multimodal visual input settings. RGB-only variants generally outperform their IR-only counterparts, consistent with the stronger alignment of RGB inputs with vision-language pretraining on natural images. Incorporating IR information improves grounding performance: MV-HiVG-B and TCFDistill exceed RGB-only HiVG-B by 5.73% and 10.01% on average across the 12 test settings, respectively. TCFDistill further outperforms RGBT-VGNet in all 12 settings, with an average gain of 2.38%. The gains average 1.98% on Test-B and 3.64% on Test-C, supporting its effectiveness under low illumination and for small targets. On Test-C, the improvements reach 3.04%, 3.35%, and 4.54% on RefFLIR, RefM^{3}FD, and RefMFAD, respectively. These results suggest that teacher-guided IR refinement and cross-modal spectral modulation enable more effective use of complementary infrared cues, improving localization when RGB observations provide limited target information.

4.3 Ablation Study

Effect of Distillation Targets. Table 2 compares DFSD with direct feature regression and frequency-decomposed feature regression. DFSD increases the mean accuracy across the four test splits from 77.06% and 76.50% to 78.80%, respectively. It also raises Test-C accuracy from 55.41% and 54.76% to 58.87%, while achieving the best results on Val, Test, Test-B, and Test-C. Although direct feature regression is slightly better on Test-A, these results favor DFSD's target-centric relational supervision over both feature-regression baselines, particularly for challenging small-target grounding.

Effect of Refinement Modality. Table 3 compares RGB-only, joint RGB–IR, and IR-only refinement. IR-only refinement performs best on Val, Test, Test-B, and Test-C, exceeding RGB-only and joint refinement by 0.43% and 0.56% on Test, with larger gains of 1.43% and 0.87% on Test-C. These results suggest that teacher-guided IR refinement compensates for the VLM's limited infrared modeling capability, enabling more effective use of complementary infrared cues for grounding.

Effect of Each Component. The ablation baseline uses our 120-epoch training protocol, whereas Table 1 reports MV-HiVG-B results from RGBT-GroundBench (see Appendix A). Table 4 evaluates the contributions of SARM and DFSD on RefM^{3}FD. SARM raises the mean accuracy across the four test splits from 76.49% to 77.87%, highlighting the benefit of local RGB–IR spectral modeling. DFSD achieves a mean accuracy of 77.64% when used alone, with gains of 1.56% and 0.92% on

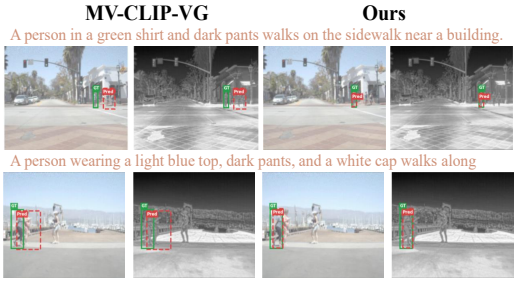


Figure 4: Qualitative comparison of RGB-T visual grounding results.

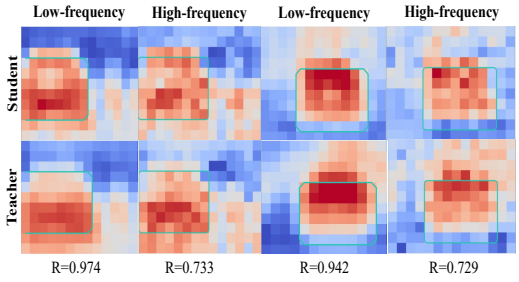


Figure 5: Visualization of the spatially structured distillation targets in DFSD.

Test-B and Test-C, respectively. Combining both components yields the highest mean accuracy of 78.80%, with improvements of 2.20% on Test and 3.25% on Test-C. These results indicate that SARM and DFSD provide complementary contributions to RGB-T grounding.

Effect of Distillation Objectives and Frequency Assignment. Tables 5 and 6 evaluate objective complementarity and frequency assignment. Combining LF-TAD and HF-SCD achieves 78.80% mean accuracy across the four test splits, exceeding either objective alone by 1.17% and 0.64%, respectively. Full-spectrum retains both relation objectives without band filtering, while Swapped reverses their frequency assignment. The proposed configuration outperforms both controls by 1.84% on Test-C, with mean gains of 1.00% and 0.89%, respectively. These results support both frequency decomposition beyond relation supervision alone and the specific pairing of low-frequency target-affinity alignment with high-frequency cross-modal correspondence alignment. SmoothL1 further exceeds MSE by 0.31% on average.

4.4 Qualitative Comparison

Figure 3 compares query-conditioned heatmaps for RGB-T visual grounding. In the illustrated example with a small referred target, TCFDistill concentrates its responses on the target, whereas RGBT-VGNet shows weaker target responses and stronger background activation. Figure 4 compares TCFDistill with MV-CLIP-VG under challenging conditions. In examples involving low illumination or small referred targets, MV-CLIP-VG is affected by weak target cues and background interference, resulting in less accurate bounding boxes. In these examples, TCFDistill produces more target-aligned localization and better separates the referred objects from nearby distractors. Figure 5 visualizes the structural relations transferred by DFSD. Student and teacher maps exhibit consistent, target-centered patterns in low-frequency target affinity and high-frequency local correspondence. Here, R denotes the Pearson correlation Pearson (1895) coefficient between paired student and teacher relation maps. These correlations suggest that the IR refinement adapter captures target-relevant teacher structures across complementary frequency bands without directly matching feature values. Additional comparisons are provided in the Appendix E.

5 Conclusion

We propose TCFDistill for RGB-T visual grounding by combining DFSD and SARM. DFSD transfers frequency-specific structural relations from the teacher model to the student branch, refining IR representations without directly matching teacher and student feature values. SARM uses local cross-modal frequency consistency to modulate shared structures and complementary responses between RGB and refined IR representations, producing spatially organized frequency tokens. Together, these components refine the VLM's IR representations and integrate complementary RGB–IR cues for query-conditioned grounding. Experiments on RefFLIR, RefM^{3}FD, and RefMFAD demonstrate improved grounding performance across diverse imaging conditions, particularly under low illumination and for small targets. Future work will extend TCFDistill to broader multimodal perception tasks.

AI USE STATEMENT

In this work, we used generative AI tools (ChatGPT, OpenAI) for language refinement and improving the clarity and organization of the manuscript. Generative AI was not used to generate experimental data, results, or scientific conclusions. All AI-assisted text was carefully reviewed and revised by the authors to ensure technical accuracy and consistency with the conducted research. We take full responsibility for the final content of this work, including any text produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on RGB-T visual grounding and does not involve human-subject experiments or the collection of personally identifiable information. All experiments are conducted on publicly available benchmark datasets and follow their corresponding licenses and usage policies. The proposed method is intended for general visual perception research, and no specific harmful or privacy-sensitive applications are considered in this study. We are not aware of any additional ethical concerns arising from the methodology or experimental evaluation presented in this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed TCFDistill framework, including DFSD and SARM, in the main paper. The experimental setup, datasets, evaluation protocols, implementation details, and training configurations are reported in the experimental section and supplementary material. Additional ablation studies and qualitative results are also provided to facilitate the verification of our findings. The source code and configuration files will be released to support reproducibility.

REFERENCES

Wei Bao, Siqi Li, Shouan Pan, Yi Xie, and Yue Gao. M^2C-EvDet: Multi-domain multi-order cross-modal knowledge distillation for event-based object detection. *IEEE Transactions on Multimedia*, 2026. doi: 10.1109/TMM.2026.3714316. Early access.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1749–1759. IEEE, 2021.

Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International conference on machine learning*, pp. 2286–2296. PMLR, 2021.

Mazin Hnewa, Alireza Rahimpour, Justin Miller, Devesh Upadhyay, and Hayder Radha. Cross modality knowledge distillation for robust pedestrian detection in low light and adverse weather conditions. In *ICASSP*, pp. 1–5, 2023.

Rui Hu, Lianghui Zhu, Yuxuan Zhang, Tianheng Cheng, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Groundingsuite: Measuring complex multi-granular pixel grounding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23105–23114. IEEE, 2025.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-2443.

Weitai Kang, Luowei Zhou, Junyi Wu, Changchang Sun, and Yan Yan. Visual grounding with attention-driven constraint balancing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1637–1645, 2025.

Yuqi Li, Chuanguang Yang, Hansheng Zeng, Zeyu Dong, Zhulin An, Yongjun Xu, Yingli Tian, and Hao Wu. Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7262–7272, 2025. doi: 10.1109/ICCV51701.2025.00682.

Fangcen Liu, Chenqiang Gao, Yaming Zhang, Junjie Guo, Jinghao Wang, and Deyu Meng. Infmae: A foundation model in the infrared modality. In *European Conference on Computer Vision*, pp. 420–437. Springer, 2024. doi: 10.1007/978-3-031-72649-1_24.

Junhong Liu, Yuan Zhang, Tao Huang, Wenchao Xu, and Renyu Yang. Distilling cross-modal knowledge via feature disentanglement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23739–23747, 2026. doi: 10.1609/aaai.v40i28.39548.

Tianshan Liu, Kin-Man Lam, Rui Zhao, and Guoping Qiu. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):315–329, 2021.

Yanfeng Liu and Lefei Zhang. Multimodal decomposed distillation with instance alignment and uncertainty compensation for thermal object detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2294–2303, 2025. doi: 10.1145/3746027.3755841.

Andong Lu, Jiacong Zhao, Chenglong Li, Yun Xiao, and Bin Luo. Breaking modality gap in RGBT tracking: Coupled knowledge distillation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. doi: 10.1145/3664647.3680878.

Fangyuan Mao, Shuo Wang, Jilin Mei, Shun Lu, Chen Min, Fuyang Liu, Xiaokun Feng, Meiqi Wu, and Yu Hu. Univ: Unified foundation model for infrared and visible modalities. *arXiv preprint arXiv:2509.15642*, 2025.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. doi: 10.1098/rspl.1895.0041.

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 *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Yasiru Ranasinghe, Elim Schenck, Florence Yellin, Shuowen Hu, Christopher Funk, and Vishal M Patel. Thermal-det: Language-guided cross-modal distillation for open-vocabulary thermal object detection. *arXiv preprint arXiv:2605.10130*, 2026.

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. *arXiv preprint arXiv:1902.09630*, 2019.

Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1181–1198, 2023.

Jingchao Wang, Wenlong Zhang, Dingjiang Huang, Hong Wang, and Yefeng Zheng. A simple and better baseline for visual grounding. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

Kunyu Wang, Guanbo Wu, Xueyang Fu, Xingbo Wang, Kean Liu, Xin Lu, Chengjie Ge, Wei Zhai, and Zheng-Jun Zha. Skyfind: A large-scale benchmark unveiling referring expression comprehension for uav. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):9859–9875, 2026. doi: 10.1109/TPAMI.2026.3681112.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *International Conference on Learning Representations*, 2024.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Ming Yan, Yaowei Wang, and Changsheng Xu. Clip-vg: Self-paced curriculum adapting of clip for visual grounding. *IEEE Transactions on Multimedia*, 26:4334–4347, 2023.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. Hivg: Hierarchical multimodal fine-grained modulation for visual grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5460–5469, 2024a.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. Oneref: Unified one-tower expression grounding and segmentation with mask referring modeling. *Advances in Neural Information Processing Systems*, 37:139854–139885, 2024b.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Toward visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3): 2749–2771, 2026.

Ruilin Yao, Shengwu Xiong, Yichen Zhao, and Yi Rong. Visual grounding with multi-modal conditional adaptation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3877–3886, 2024.

Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15481–15491. IEEE, 2022.

Maoxun Yuan, Bo Cui, Tianyi Zhao, Jiayi Wang, Shan Fu, Xue Yang, and Xingxing Wei. Unirgb-ir: A unified framework for visible-infrared semantic tasks via adapter tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2409–2418, 2025.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

Yuan Zhang, Tao Huang, Jiaming Liu, Tao Jiang, Kuan Cheng, and Shanghang Zhang. Freekd: Knowledge distillation via semantic frequency prompt. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15931–15940. IEEE, 2024.

Tianyi Zhao, Jiawen Xi, Linhui Xiao, Junnan Li, Xue Yang, Maoxun Yuan, and Xingxing Wei. Rgbt-groundbench: Visual grounding beyond rgb in complex real-world scenarios. In *European Conference on Computer Vision (ECCV)*, 2026.

Kailai Zhou, Fuqiang Yang, Shixian Wang, Bihan Wen, Chongde Zi, Linsen Chen, Qiu Shen, and Xun Cao. M-specgene: Generalized foundation model for rgbt multispectral vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7861–7872, October 2025.

APPENDIX

A Additional Experimental and Implementation Details

A.1 Training and Implementation Settings

Following the initialization protocol of RGBT-GroundBench Zhao et al. (2026), we initialize TCFDistill from the pretrained HiVG RGB-T checkpoint. The original CLIP parameters and the teacher model remain frozen, while the lightweight adapters and task-specific modules are optimized.

SARM uses overlapping 3×3 local windows with a frequency embedding dimension of 64. The reconstructed representations form a 14×14 spatial grid, yielding 196 frequency tokens. The dropout rate within SARM is set to zero.

For DFSD, we set the frequency cutoff to 0.5 and the RGB–text response temperature to 0.07. LF-TAD and HF-SCD are weighted by 1.0 and 0.1, respectively, and the overall DFSD objective is weighted by 0.05. The LF-TAD and HF-SCD relation objectives are optimized using SmoothL1 loss.

Within HF-SCD, the point-wise correspondence loss has unit weight, while the local correspondence loss is weighted by 0.2. The rank of the low-rank IR refinement adapter is set to 16. Its output is projected through a zero-initialized linear layer and selectively injected into the IR representation using an RGB–text relevance gate. All remaining module settings are shared across the three datasets.

The teacher model provides supervision only during training and is removed at inference. We verified that the teacher pretraining data contain no images from the validation or test splits of RefFLIR, RefM³FD, and RefMFAD.

Baseline Training Protocol. The MV-HiVG-B results in Table 1 are reported from RGBT-GroundBench Zhao et al. (2026). All experiments conducted for TCFDistill, including its ablations and the retrained HiVG baseline, use a single-stage training protocol of 120 epochs. HiVG's existing adaptation modules are jointly optimized from the start, while the original CLIP weights remain frozen. This replaces the warm-up and three successive adaptation stages of the original HiVG schedule. The retrained baseline excludes SARM and DFSD and provides a reference for assessing their contributions under the same training schedule.

A.2 Evaluation Subset Definitions

Following the official RGBT-GroundBench protocol Zhao et al. (2026), Test-A contains normal-size targets under normal- or strong-light conditions, Test-B focuses on nighttime scenes under weak- or very-weak-light conditions, and Test-C contains small-size targets. The numbers of instances in Test-A, Test-B, and Test-C are 837, 640, and 968 for RefFLIR; 1,232, 1,094, and 1,848 for RefM^{3}FD; and 789, 2,452, and 2,544 for RefMFAD, respectively. These condition-specific subsets are evaluated independently using Acc@0.5. Accordingly, Test-B and Test-C are used to assess grounding under low illumination and for small targets, respectively.

A.3 DFSD Implementation Details

Gradient flow and parameter updates. We apply stop-gradient to the LF-TAD query weights, the student RGB features used by HF-SCD, and the teacher-derived targets, directing distillation gradients to the IR refinement adapter. The grounding and auxiliary losses also optimize the trainable student modules, including the IR refinement adapter.

Feature centering and radial masks. Given patch features $\boldsymbol{X} \in \mathbb{R}^{N \times d}$, where $N = H \times W$, we reshape them into an $H \times W$ spatial grid. For each channel, we subtract the mean over valid spatial positions and set padded positions to zero, obtaining the centered grid $\widetilde{\boldsymbol{X}}$.

For the centered Fourier spectrum, we define the normalized radius as

$$r_{h,w} = \frac{1}{\sqrt{2}} \sqrt{\left(\frac{h - \lfloor H/2 \rfloor}{\max(H/2, 1)} \right)^2 + \left(\frac{w - \lfloor W/2 \rfloor}{\max(W/2, 1)} \right)^2}. \tag{19}$$

The complementary low- and high-frequency masks are

$$M_l(h, w) = \mathbb{I}[r_{h,w} \leq c], \qquad M_h(h, w) = 1 - M_l(h, w), \tag{20}$$

where $c = 0.5$, with frequencies on the cutoff boundary assigned to the low-frequency band. Both masks are shared across channels.

Frequency decomposition. For $b \in \{l, h\}$, the band-specific features are computed as

$$\boldsymbol{X}^b = \mathcal{R}^{-1}\Big(\mathrm{real}\Big(\mathcal{F}^{-1}\Big[\mathrm{ifftshift}\Big(M_b \odot \mathrm{fftshift}\Big(\mathcal{F}(\widetilde{\boldsymbol{X}})\Big)\Big)\Big]\Big)\Big), \tag{21}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the channel-wise two-dimensional FFT and inverse FFT with orthonormal normalization. The shift operations align the spectrum with the centered radial masks. We retain the real part of the inverse transform, and $\mathcal{R}^{-1}$ restores the token sequence, yielding $\boldsymbol{X}^b \in \mathbb{R}^{N \times d}$ with the original spatial ordering.

Target-region masking. We apply the same geometric transformations to the image and its binary ground-truth box mask, with padded pixels assigned zero. The mask is resized to the 14×14 patch grid using nearest-neighbor interpolation, and positions with values above 0.5 form Ω. If Ω is empty, we select the grid cell containing the ground-truth box center, with its indices clamped to the grid bounds. Both HF-SCD losses are averaged over Ω for each image and then across the batch. Local correspondence averages exclude padded positions and neighbors outside the grid.

B Analysis of Cross-Modal Structural Consistency

We quantify the structural consistency illustrated in Fig. 1(a) by measuring how well cross-modal RGB–IR relations preserve the spatial similarity structure of RGB representations. The analysis uses the final visual patch representations from the VLM (e.g., CLIP) and the pretrained RGB–IR teacher model (e.g., UNIV Mao et al. (2025)). Both models process the same image pairs, and the relations are computed independently within their respective representation spaces.

Let $\mathcal{V}$ denote the set of non-padded visual patch positions. For each RGB anchor at position $i \in \mathcal{V}$, we construct two relation vectors:

$$\boldsymbol{a}_i = \left[\cos\left(\boldsymbol{x}_i^{rgb}, \boldsymbol{x}_j^{rgb}\right)\right]_{j \in \mathcal{V}}, \qquad \boldsymbol{b}_i = \left[\cos\left(\boldsymbol{x}_i^{rgb}, \boldsymbol{x}_j^{ir}\right)\right]_{j \in \mathcal{V}}. \tag{22}$$

Here, $\boldsymbol{x}_i^{rgb}$ denotes the RGB feature at the anchor position, while $\boldsymbol{x}_j^{rgb}$ and $\boldsymbol{x}_j^{ir}$ denote the RGB and IR features at position j, respectively. Thus, $\boldsymbol{a}_i$ encodes intra-modal RGB relations, whereas $\boldsymbol{b}_i$ encodes cross-modal RGB–IR relations.

To compare relative spatial structures independently of absolute similarity levels, we mean-center both relation vectors:

$$\bar{a}_i = \mathrm{mean}(\boldsymbol{a}_i), \qquad \bar{b}_i = \mathrm{mean}(\boldsymbol{b}_i), \qquad \widetilde{\boldsymbol{a}}_i = \boldsymbol{a}_i - \bar{a}_i, \qquad \widetilde{\boldsymbol{b}}_i = \boldsymbol{b}_i - \bar{b}_i. \tag{23}$$

The structural consistency score for an image pair is computed by averaging the relation agreement over the valid anchor positions:

$$s_{\mathrm{str}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \cos\left(\widetilde{\boldsymbol{a}}_i, \widetilde{\boldsymbol{b}}_i\right). \tag{24}$$

A higher s_{str} indicates better preservation of the RGB spatial similarity structure in the cross-modal RGB–IR relations.

We evaluate all unique RGB–IR image pairs in the validation splits of RefFLIR, RefM^{3}FD, and RefMFAD, covering diverse target objects and imaging conditions. For images associated with

multiple referring expressions, each RGB–IR image pair is counted only once. In Fig. 1(a), the bars report the mean image-level structural consistency score, and the error bars denote the standard deviation across images.

UNIV achieves mean scores of 0.966, 0.967, and 0.961, compared with 0.708, 0.537, and 0.538 for CLIP. These results indicate that the pretrained RGB–IR teacher provides more consistent cross-modal structural relations, supporting its use for structural distillation. The scores measure relational consistency rather than grounding accuracy or frequency-specific distortion.

C Training Objectives

We train TCFDistill using the grounding losses inherited from HiVG and the proposed DFSD objective. Given a predicted normalized bounding box $\widehat{B} \in [0,1]^4$ and its ground truth $B \in [0,1]^4$, the localization loss combines ℓ_1 regression and generalized IoU loss Rezatofighi et al. (2019):

$$\mathcal{L}_{\text{loc}} = \lambda_{\ell_1} \left\| \widehat{B} - B \right\|_1 + \lambda_{\text{giou}} \left[1 - \text{GIoU}(\widehat{B}, B) \right]. \tag{25}$$

The two terms supervise bounding-box coordinates and spatial overlap, respectively.

Table 7: Component-wise ablation of TCFDistill on RefFLIR, RefM^{3}FD, and RefMFAD. Here, † indicates that the complete refinement adapter, projection, reinjection path, and SARM are retained, while only the DFSD distillation loss is disabled by setting $\lambda_{\text{kd}} = 0$. The best and second-best results in each column are highlighted in bold and underlined, respectively.

Variant	SARM	DFSD	RefFLIR					RefM^{3}FD					RefMFAD				
			Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
A: Baseline			76.64	72.14	89.26	68.12	51.41	72.02	75.13	94.23	80.98	55.62	70.78	69.13	91.13	66.23	53.96
B: + SARM	✓		**78.12**	72.86	89.26	67.50	53.24	75.00	76.66	95.26	82.35	57.19	<u>71.97</u>	69.29	90.88	66.72	54.07
C: + DFSD		✓	**78.12**	73.19	89.85	68.59	53.34	74.40	76.11	<u>95.37</u>	82.54	56.54	71.65	69.81	**91.77**	66.84	54.66
D†: Full architecture w/o KD	✓	✓†	77.08	<u>74.21</u>	<u>90.12</u>	<u>69.12</u>	<u>54.17</u>	<u>76.54</u>	<u>76.81</u>	95.21	<u>82.97</u>	<u>57.82</u>	71.88	<u>70.21</u>	91.54	<u>67.07</u>	<u>55.11</u>
E: TCFDistill (Full)	✓	✓	<u>77.13</u>	**75.00**	**91.88**	**69.84**	**55.57**	**77.38**	**77.33**	**95.45**	**83.54**	**58.87**	**72.05**	**70.40**	<u>91.64</u>	**67.33**	**55.72**

DFSD supervises the IR refinement adapter through complementary low- and high-frequency structural objectives:

$$\mathcal{L}_{\text{DFSD}} = \lambda_L \mathcal{L}_{\text{LF-TAD}} + \lambda_H \mathcal{L}_{\text{HF-SCD}}. \tag{26}$$

The overall training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{aux}} + \lambda_{\text{kd}} \mathcal{L}_{\text{DFSD}}, \tag{27}$$

where $\mathcal{L}_{\text{aux}}$ comprises the contrastive, patch-level alignment, and auxiliary segmentation losses inherited from HiVG.

We set $\lambda_{\ell_1} = \lambda_{\text{giou}} = 2$, $\lambda_L = 1$, $\lambda_H = 0.1$, and $\lambda_{\text{kd}} = 0.05$. The teacher model remains frozen and provides supervision only during training.

D Additional Ablation Studies

D.1 Additional Component-wise Ablation Results.

Table 7 verifies the contributions of SARM, DFSD, and DFSD teacher supervision across all three datasets. For the SARM-only variant, we directly use the original IR representation, i.e., $\boldsymbol{Z}^{ir+} = \boldsymbol{Z}^{ir}$, as the IR input to SARM, since the DFSD-based IR refinement is disabled. The SARM-only and DFSD-only variants improve the mean accuracy across the 12 test settings by 0.66% and 0.94%, respectively. SARM achieves its largest dataset-level gain of 1.38% on RefM^{3}FD, while DFSD improves Test-B and Test-C by average gains of 0.88% and 1.18%, respectively.

The full-architecture no-distillation control retains the refinement adapter, projection, reinjection path, and SARM, while setting $\lambda_{\text{kd}} = 0$. It achieves a mean accuracy of 73.70%, compared with 74.38% for the full model. The resulting 0.68% gain from enabling DFSD teacher supervision indicates that the improvement cannot be explained solely by the capacity of the newly introduced refinement and modulation modules.

Table 8: Ablation of the low- and high-frequency objectives in DFSD. The best and second-best results in each column are highlighted in bold and underlined, respectively.

Variant	LF-TAD	HF-SCD	Regression	RefFLIR					RefM^{3}FD					RefMFAD				
				Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
LF-TAD only	✓		SmoothL1	**78.61**	73.09	90.45	68.75	53.14	76.19	76.11	95.21	82.54	56.65	71.89	69.62	90.88	66.68	55.01
HF-SCD only		✓	SmoothL1	77.79	74.09	90.81	**70.15**	54.76	**77.97**	76.85	95.12	82.17	58.49	**72.29**	69.55	91.13	**67.33**	54.43
TCFDistill with MSE	✓	✓	MSE	77.96	73.43	90.33	68.12	53.44	75.59	77.11	**95.86**	82.44	58.54	**72.29**	69.62	91.13	66.55	54.50
DFSD (Full)	✓	✓	SmoothL1	77.13	**75.00**	**91.88**	69.84	**55.57**	77.38	**77.33**	95.45	**83.54**	**58.87**	72.05	**70.40**	**91.64**	**67.33**	**55.72**

Combining SARM and DFSD raises the overall mean accuracy from 72.28% to 74.38%, exceeding the SARM-only and DFSD-only variants by 1.44% and 1.16%, respectively. The full model also improves over the no-distillation control by 0.68%. On Test-C, the gain over the baseline reaches 3.06%, highlighting the complementary contributions of local spectral modulation and frequency-specific structural alignment.

Overall, TCFDistill improves RGB-T visual grounding by learning query-conditioned, spectrally aligned RGB–IR representations that better exploit complementary infrared cues, particularly under low illumination and for small referred targets.

D.2 Additional Ablation of Dual-Frequency Objectives

Table 8 examines the contributions of low- and high-frequency structural alignment. Under SmoothL1-based relation alignment, enabling both objectives yields higher test accuracy than either objective alone across all three datasets. On RefFLIR, LF-only and HF-only alignment achieve 73.09% and 74.09%, respectively, while their combination reaches 75.00%. On RefM^{3}FD Test-C, joint alignment achieves 58.87%, compared with 56.65% for LF-only and 58.49% for HF-only alignment. These results demonstrate the complementary contributions of query-conditioned low-frequency target-affinity relations and high-frequency RGB–IR correspondences. Combining coarse-scale target structure with fine-grained correspondence relations defined at matched positions and within local neighborhoods provides more effective structural supervision for the IR refinement adapter than either frequency objective alone.

With both objectives enabled, SmoothL1 also yields higher test accuracy than MSE across all three datasets. For example, accuracy increases from 73.43% to 75.00% on RefFLIR and from 69.62% to 70.40% on RefMFAD. Together, these results support the DFSD design: SmoothL1-based alignment of complementary low- and high-frequency structural relations improves grounding performance.

D.3 Additional Ablation of Local Spectral Interaction in SARM.

We investigate the contributions of local frequency decomposition and cross-modal spectral interaction through two controlled ablations of SARM. The *spatial-neighborhood* variant replaces DCT/IDCT with direct processing and overlap reconstruction of 3×3 spatial windows. The shared projection, cosine consistency, two-stream consistency-conditioned modulation, fusion adapter, and output tokenization remain unchanged, providing a parameter-matched comparison between spatial and frequency modeling.

The IR-only local DCT variant retains DCT/IDCT but removes RGB features and cross-modal consistency modulation. Specifically, the RGB spectral stream is set to zero, while the IR spectral stream is directly retained, preserving the fusion-adapter architecture and parameter count. This comparison evaluates whether RGB–IR spectral interaction provides additional benefits over IR frequency encoding alone.

Table 9 shows that full SARM achieves 74.38% mean accuracy across the 12 test splits, exceeding the spatial-neighborhood and IR-only variants by 0.98% and 1.23%, respectively. The corresponding gains on Test-C are 1.01% and 1.33%, averaged across the three datasets. Overall, these ablations show that SARM improves grounding by combining local frequency decomposition with RGB–IR spectral interaction. SARM uses cross-modal frequency consistency to modulate paired RGB and IR spectral responses, followed by spatial reconstruction to produce frequency tokens.

Table 9: Ablation of local spectral interaction and RGB–IR coupling in SARM. The spatial-neighborhood variant replaces local DCT/IDCT with parameter-matched spatial-neighborhood modeling, whereas the IR-only variant removes RGB features and cross-modal interaction. DFSD and all other settings remain unchanged. The best and second-best results in each column are highlighted in bold and underlined, respectively.

Variant	Local DCT	RGB–IR interaction	RefFLIR					RefM^{3}FD					RefMFAD				
			Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
Spatial-neighborhood		✓	<u>78.12</u>	<u>73.43</u>	<u>89.73</u>	<u>69.06</u>	<u>54.66</u>	**77.97**	<u>75.69</u>	<u>95.29</u>	<u>83.08</u>	<u>58.00</u>	<u>72.21</u>	69.50	<u>91.01</u>	<u>66.92</u>	54.47
IR-only local DCT	✓		**78.78**	72.67	89.26	68.59	52.94	75.59	75.48	95.21	82.99	57.95	**72.29**	<u>69.79</u>	90.88	66.76	<u>55.29</u>
SARM (Full)	✓	✓	77.13	**75.00**	**91.88**	**69.84**	**55.57**	<u>77.38</u>	**77.33**	**95.45**	**83.54**	**58.87**	72.05	**70.40**	**91.64**	**67.33**	**55.72**

Table 10: Ablation of the point-wise and local correspondence losses in HF-SCD. The best and second-best results in each column are highlighted in bold and underlined, respectively.

Variant	$\mathcal{L}_{\text{point}}$	$\mathcal{L}_{\text{local}}$	RefFLIR					RefM^{3}FD					RefMFAD				
			Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
Point-wise loss only	✓		76.80	<u>73.62</u>	<u>90.21</u>	<u>69.60</u>	<u>53.85</u>	76.78	76.50	**95.77**	<u>82.54</u>	57.30	<u>72.13</u>	70.16	<u>90.88</u>	<u>67.21</u>	**55.76**
Local loss only		✓	**77.30**	72.76	89.73	67.50	53.24	**78.57**	<u>76.64</u>	<u>95.45</u>	<u>82.54</u>	<u>57.57</u>	**72.37**	<u>70.38</u>	90.37	66.94	55.50
HF-SCD (Full)	✓	✓	<u>77.13</u>	**75.00**	**91.88**	**69.84**	**55.57**	<u>77.38</u>	**77.33**	<u>95.45</u>	**83.54**	**58.87**	72.05	**70.40**	**91.64**	**67.33**	<u>55.72</u>

D.4 Additional Ablation Studies of DFSD

Ablation of Point-wise and Local Correspondence Losses in HF-SCD. Table 10 evaluates the point-wise and local correspondence objectives in HF-SCD. The point-wise loss achieves a higher mean accuracy than the local loss across the 12 test settings (73.62% vs. 73.22%), indicating the importance of preserving same-position RGB-IR associations. The local loss remains competitive on RefM^{3}FD, suggesting that neighborhood relations provide complementary structural information. Combining both losses yields the highest mean accuracy of 74.38%, outperforming the point-wise-only and local-only variants by 0.76% and 1.16%, respectively. The gains on RefFLIR and the challenging Test-B and Test-C subsets further demonstrate the complementarity between direct patch correspondence and local structural context.

Additional Ablation on Distillation Targets. We compare DFSD with two feature-regression variants. Direct feature regression matches the student and teacher IR patch features without frequency decomposition. Frequency-decomposed feature regression instead matches their low- and high-frequency features separately using uniform spatial weighting. In contrast, target-centric DFSD distills query-conditioned low-frequency target-affinity relations and target-region-focused high-frequency RGB–IR correspondences at matched positions and within local neighborhoods, rather than directly matching feature values.

As shown in Table 11, the two feature-regression variants achieve mean accuracies of 73.44% and 73.25% across the 12 test splits, respectively. DFSD achieves 74.38%, improving the mean accuracy by 0.94% and 1.13%. The gains further increase to 3.46% and 4.11% on RefM^{3}FD Test-C. These results show that relation-based dual-frequency distillation transfers target-relevant cross-modal structural supervision more effectively than the evaluated feature-regression objectives, particularly for small-target grounding.

Additional Ablation on Dual-Frequency Distillation. Table 12 examines the contributions of low- and high-frequency structural distillation. Using SmoothL1 for relation distillation, enabling both objectives yields higher Test accuracy than either objective alone on all three datasets. On RefFLIR, LF-only and HF-only distillation achieve 73.09% and 74.09%, respectively, while their combination reaches 75.00%. On RefM^{3}FD Test-C, joint distillation achieves 58.87%, compared with 56.65% for LF-only and 58.49% for HF-only distillation. These results show the complementary contributions of query-conditioned low-frequency target-affinity relations and target-region-focused high-frequency RGB–IR correspondences. Combining coarse-scale target structure with fine-grained correspondence relations defined at matched positions and within local neighborhoods provides more effective structural supervision for the IR refinement adapter than either frequency objective alone.

Table 11: Comparison of distillation targets in DFSD. The two feature-regression variants replace the structural objectives with direct feature regression or frequency-decomposed feature regression. The best and second-best results in each column are highlighted in bold and underlined, respectively.

Distillation target	RefFLIR					RefM^{3}FD					RefMFAD				
	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
Direct feature regression	77.30	74.33	91.28	69.84	54.86	76.19	75.52	**95.61**	81.71	55.41	**72.53**	69.65	90.75	**67.41**	54.86
Frequency-decomposed feature regression	**77.79**	74.04	90.21	**70.31**	54.76	76.19	74.94	95.21	81.07	54.76	72.45	70.24	90.63	67.25	55.60
DFSD (Full)	77.13	**75.00**	**91.88**	69.84	**55.57**	**77.38**	**77.33**	95.45	**83.54**	**58.87**	72.05	**70.40**	**91.64**	67.33	**55.72**

Table 12: Ablation of DFSD's dual-frequency objectives and regression loss. The best and second-best results in each column are highlighted in bold and underlined, respectively.

Variant	LF	HF	Loss	RefFLIR					RefM^{3}FD					RefMFAD				
				Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
LF only	✓		SmoothL1	**78.61**	73.09	90.45	68.75	53.14	76.19	76.11	95.21	82.54	56.65	71.89	69.62	90.88	66.68	55.01
HF only		✓	SmoothL1	77.79	74.09	90.81	**70.15**	54.76	**77.97**	76.85	95.12	82.17	58.49	**72.29**	69.55	91.13	**67.33**	54.43
LF + HF	✓	✓	MSE	77.96	73.43	90.33	68.12	53.44	75.59	77.11	**95.86**	82.44	58.54	**72.29**	69.62	91.13	66.55	54.50
LF + HF	✓	✓	SmoothL1	77.13	**75.00**	**91.88**	69.84	**55.57**	77.38	**77.33**	95.45	**83.54**	**58.87**	72.05	**70.40**	**91.64**	**67.33**	**55.72**

With both objectives enabled, SmoothL1 also yields higher Test accuracy than MSE on all three datasets. For example, accuracy increases from 73.43% to 75.00% on RefFLIR and from 69.62% to 70.40% on RefMFAD. Together, these results show that target-centric DFSD with SmoothL1 distillation of complementary low- and high-frequency structural relations improves grounding performance.

Additional Ablation of Frequency Assignment. Table 13 evaluates the role of frequency decomposition and the assignment of distillation objectives. Full-spectrum applies both TAD and SCD to features without band filtering, while Swapped applies TAD to high-frequency features and SCD to low-frequency features. Both retain the corresponding relation definitions and loss weights.

Across the 12 test settings, TCFDistill achieves a mean accuracy of 74.38%, outperforming Full-spectrum and Swapped by 0.83% and 1.32%, respectively. It also achieves higher Test and Test-C accuracy than both variants on all three datasets. The improvement over Full-spectrum supports the value of explicitly separating frequency bands beyond applying relation supervision alone. The advantage over Swapped further supports the specific pairing of low-frequency target-affinity alignment with high-frequency cross-modal correspondence alignment. These results provide evidence that both frequency decomposition and the assignment of structural objectives contribute to DFSD's effectiveness.

D.5 Additional Analysis of Query-Conditioned Correction Reinjection

During training, DFSD supervises an auxiliary IR refinement branch using query-conditioned low-frequency target-affinity alignment and high-frequency RGB–IR correspondence alignment. The auxiliary branch applies a residual adapter to the gradient-detached IR representation, producing $\boldsymbol{Z}^{aux}$. Let $\boldsymbol{Z}^{aux} = [\boldsymbol{z}_i^{aux}]_{i=1}^{N}$ and $\boldsymbol{Z}^{ir} = [\boldsymbol{z}_i^{ir}]_{i=1}^{N}$ denote the auxiliary and original IR token sequences, respectively. Its projected output is selectively re-injected into the original IR representation:

$$\boldsymbol{z}_i^{ir+} = \boldsymbol{z}_i^{ir} + g_i \phi_0(\boldsymbol{z}_i^{aux}), \tag{28}$$

where ϕ_0 is a bias-free linear projection initialized with zero weights. Here, g_i is the scalar RGB–text relevance gate at the i-th spatial position, and the gate sequence is

$$\boldsymbol{G} = G(\boldsymbol{Z}^{rgb}, \boldsymbol{Z}^{q}) = [g_i]_{i=1}^{N}.$$

The gate is computed from the same patch-wise RGB–text cosine similarities used in LF-TAD. Let $\bar{\boldsymbol{z}}^q = \mathrm{Pool}(\boldsymbol{Z}^q)$ denote the global query representation defined in Eq. 6, and let $\mathcal{V}$ denote the set of non-padded visual positions. We compute

$$s_i = \cos\left(\boldsymbol{z}_i^{rgb}, \bar{\boldsymbol{z}}^q\right), \qquad \mu_s = \frac{1}{|\mathcal{V}|}\sum_{j\in\mathcal{V}} s_j, \tag{29}$$

Table 13: Ablation of frequency assignment in DFSD across three datasets. The TAD and SCD columns specify the frequency band used by each objective; "–" denotes an inactive objective. The best and second-best results are highlighted in bold and underlined, respectively.

Variant	TAD	SCD	RefFLIR					RefM^{3}FD					RefMFAD				
			Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
HF-TAD only	High	–	77.79	72.38	89.73	67.34	51.82	75.00	76.11	**95.61**	83.27	56.38	**72.37**	69.58	91.26	66.84	54.35
LF-SCD only	–	Low	**78.12**	73.52	88.30	69.18	54.25	76.78	76.00	95.45	81.99	56.54	71.97	69.50	90.50	66.80	54.58
Full-spectrum	Full	Full	76.97	73.71	89.26	69.53	54.66	74.40	76.32	95.21	82.63	57.03	71.17	70.00	**92.53**	66.96	54.82
Swapped	High	Low	76.64	72.76	89.97	67.50	51.52	76.78	77.03	95.04	82.54	57.03	71.49	69.91	91.51	67.08	54.90
TCFDistill (Full)	Low	High	77.13	**75.00**	**91.88**	**69.84**	**55.57**	**77.38**	**77.33**	95.45	**83.54**	**58.87**	72.05	**70.40**	91.64	**67.33**	**55.72**

Table 14: Ablation of correction reinjection in TCFDistill. All variants use RGB–text-guided KD and differ only in the reinjection gate. The best and second-best results in each column are highlighted in bold and underlined, respectively.

Variant	Reinjection gate	RefFLIR					RefM^{3}FD					RefMFAD				
		Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
Fixed addition	$g_i = 1$	76.64	74.14	90.45	**70.00**	54.05	75.00	76.48	94.88	83.08	57.68	71.25	69.76	90.37	66.88	54.98
No reinjection	$g_i = 0$	77.13	72.24	88.54	67.96	52.12	76.19	76.50	**95.61**	82.72	57.25	**72.37**	69.67	91.13	66.80	54.78
Uniform half injection	$g_i = 0.5$	**77.30**	72.62	90.57	69.68	50.81	75.00	75.55	95.21	82.35	55.68	**72.37**	69.72	90.50	67.00	54.98
TCFDistill (Full)	RGB–text g_i	77.13	**75.00**	**91.88**	69.84	**55.57**	**77.38**	**77.33**	95.45	**83.54**	**58.87**	72.05	**70.40**	**91.64**	**67.33**	**55.72**

and define

$$g_i = \begin{cases} \sigma(s_i - \mu_s), & i \in \mathcal{V}, \\ 0, & i \notin \mathcal{V}. \end{cases} \tag{30}$$

Unlike the temperature-scaled softmax used to obtain the LF-TAD weights p_i, the gate applies a sigmoid to the mean-centered similarity scores. The scalar gates $\{g_i\}$ form the matrix gate $\boldsymbol{G}$ used in the main text after being broadcast along the feature dimension. The gate is parameter-free and detached from gradient computation when applied to $\phi_0(\boldsymbol{Z}^{aux})$. Consequently, patches with above-average RGB–text similarity receive stronger correction injection, while padded positions receive no correction.

We compare three fixed-gate controls with the full query-conditioned strategy. All variants retain the same RGB–text-guided target weighting in DFSD and differ only in the correction reinjection gate. Fixed addition sets $g_i = 1$ at every valid spatial position, applying the correction without spatial selection. No reinjection sets $g_i = 0$, preventing the learned correction from entering the grounding branch. Uniform half injection sets $g_i = 0.5$, applying the correction uniformly with reduced strength. The full TCFDistill model uses the spatially varying gate defined in Eq. 30.

As shown in Table 14, fixed addition, no reinjection, and uniform half injection achieve mean accuracies of 73.56%, 72.94%, and 72.89% across the 12 test splits, respectively. The full TCFDistill model achieves the highest mean accuracy of 74.38%, outperforming these variants by 0.82%, 1.44%, and 1.49%, respectively. The improvement over no reinjection shows the benefit of incorporating the DFSD-supervised correction into the grounding representation. Its consistent advantage over the fixed-gate variants further shows that target-centric, spatially selective reinjection is more effective than uniform-strength injection.

D.6 Additional Analysis of Query Guidance

We analyze the two uses of query guidance in TCFDistill: query-conditioned target weighting in LF-TAD and query-conditioned correction reinjection. The query-agnostic variant uses uniform LF-TAD weighting and a fixed reinjection gate of $g_i = 0.5$, where g_i controls the correction strength at spatial position i. The LF-TAD-guided-only variant applies RGB–text guidance to LF-TAD while retaining the fixed gate, whereas the injection-guided-only variant uses uniform LF-TAD weighting with the RGB–text relevance gate. The full model applies query guidance to both operations.

As shown in Table 15, the query-agnostic, LF-TAD-guided-only, and injection-guided-only variants achieve mean accuracies of 73.31%, 72.89%, and 73.69% across the 12 test splits, respectively. The full TCFDistill model achieves the highest mean accuracy of 74.38%, exceeding these variants by

Table 15: Ablation of query guidance in LF-TAD and correction reinjection. Uniform LF-TAD weighting assigns equal weights to spatial positions. The LF-TAD spatial-shuffle variant preserves the query-weight distribution but randomly permutes it across valid spatial positions, while retaining the RGB–text reinjection gate. Here, $g_i = 0.5$ denotes uniform correction reinjection. The best and second-best results in each column are highlighted in bold and underlined, respectively.

Variant	LF-TAD target weighting	Reinjection gate	RefFLIR					RefM^{3}FD					RefMFAD				
			Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
Query-agnostic	Uniform	$g_i = 0.5$	**78.45**	73.05	89.14	68.59	54.15	76.19	77.03	95.43	81.80	58.53	71.97	69.65	90.50	66.68	55.13
LF-TAD-guided only	RGB–text	$g_i = 0.5$	77.30	72.62	90.57	69.68	50.81	75.00	75.55	95.21	82.35	55.68	**72.37**	69.72	90.50	67.00	54.98
LF-TAD spatial-shuffle	Shuffled RGB–text	RGB–text	77.12	73.47	89.14	69.25	54.25	75.59	76.16	94.61	82.54	56.60	71.97	70.12	91.01	67.08	54.17
Injection-guided only	Uniform	RGB–text	76.80	74.38	90.93	**70.93**	54.76	76.19	76.35	95.41	81.80	57.19	72.13	69.88	90.37	67.00	55.25
TCFDistill (Full)	RGB–text	RGB–text	77.13	**75.00**	**91.88**	69.84	**55.57**	**77.38**	**77.33**	**95.45**	**83.54**	**58.87**	72.05	**70.40**	**91.64**	**67.33**	**55.72**

1.07 %, 1.49%, and 0.69%, respectively. Under a fixed reinjection gate, LF-TAD guidance alone does not improve the mean accuracy, while query-guided reinjection provides a gain of 0.38% over the query-agnostic variant. Combining the two forms of guidance yields the best overall result, supporting their joint use for target-focused distillation and correction reinjection.

Uniform LF-TAD weighting sets $p_i = 1/|\mathcal{V}|$ for $i \in \mathcal{V}$, where p_i is the weight used to construct the low-frequency target prototype and $\mathcal{V}$ denotes the set of valid, non-padded spatial positions. Because frequency decomposition follows spatial mean-centering, the uniform prototype can collapse to zero when all positions are valid. Excluding padded positions need not preserve this cancellation because frequency filtering mixes spatial locations. Thus, uniform weighting alone does not isolate the contribution of query-conditioned spatial weighting.

To assess whether the spatial arrangement of query weights matters, we retain the full model's RGB–text reinjection gate and randomly shuffle the LF-TAD query weights among valid spatial positions within each image. This preserves the weight values and their sum while disrupting their correspondence with image locations. The spatial-shuffle variant achieves a mean accuracy of 73.20%, compared with 74.38% for the full model. The full model outperforms this control on all 12 test splits, with a mean accuracy gain of 1.18%. These results support the importance of spatially aligned query weighting in LF-TAD, beyond non-uniform weighting alone, when paired with query-guided correction reinjection.

D.7 Additional Analysis of Teacher Selection.

We further investigate the influence of teacher selection by replacing UNIV with M-SpecGene Zhou et al. (2025) and InfMAE Liu et al. (2024) while retaining the same student framework. As shown in Table 16, UNIV achieves the highest mean accuracy of 74.38% across the 12 test splits, outperforming M-SpecGene and InfMAE by 0.75% and 1.21%, respectively. Its advantage is strongest on RefFLIR and remains consistent on RefM^{3}FD and RefMFAD, indicating more reliable transfer across different imaging conditions.

M-SpecGene remains competitive on RefM^{3}FD and achieves a higher accuracy on Test-A, while UNIV performs better on Test, Test-B, and Test-C. However, it performs less consistently on RefFLIR and RefMFAD. InfMAE achieves the highest RefMFAD validation accuracy but shows lower performance on the corresponding test splits. These results indicate that thermal representations provide useful structural supervision, whereas the paired RGB–IR representations of UNIV are better aligned with the cross-modal structural relations distilled by DFSD. Overall, teacher selection affects both average accuracy and cross-dataset stability, and UNIV provides the most reliable frequency-specific structural supervision among the evaluated teachers.

D.8 Additional Ablation on the Refinement Modality.

Table 17 compares RGB-only, IR-only, and joint RGB–IR refinement. IR-only refinement achieves the highest Test accuracy across all three datasets. On RefFLIR, it surpasses RGB-only and joint refinement by 1.72% and 1.95% on Test, respectively, and exceeds RGB-only refinement by 1.79% and 2.33% on Test-A and Test-C. On RefM^{3}FD and RefMFAD, IR-only refinement outperforms both alternatives on Test, Test-B, and Test-C. In particular, its Test-C gains on RefM^{3}FD reach 1.43% and 0.87% over RGB-only and joint refinement, respectively. These results support directing

Table 16: Comparison of different teacher models on RefFLIR, RefM^{3}FD, and RefMFAD. The best and second-best results are highlighted in bold and underlined, respectively.

Teacher	RefFLIR					RefM^{3}FD					RefMFAD				
	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
M-SpecGene Zhou et al. (2025)	76.64	73.14	90.09	69.21	53.34	74.40	77.06	**96.02**	82.26	58.60	71.49	69.98	91.39	**67.33**	55.13
InfMAE Liu et al. (2024)	76.97	72.81	89.73	68.59	52.83	74.40	76.64	95.12	82.72	57.95	**73.16**	69.55	90.88	66.72	54.47
UNIV Mao et al. (2025)	**77.13**	**75.00**	**91.88**	**69.84**	**55.57**	**77.38**	**77.33**	95.45	**83.54**	**58.87**	72.05	**70.40**	**91.64**	**67.33**	**55.72**

Table 17: Additional Ablation on the refinement modality across RefFLIR, RefM^{3}FD, and RefMFAD. IR-only refinement corresponds to the full TCFDistill model. The best and second-best results are highlighted in bold and underlined, respectively.

Refined modality	RefFLIR					RefM^{3}FD					RefMFAD				
	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
RGB only	**78.45**	73.28	90.09	**70.00**	53.24	76.19	76.90	95.37	82.81	57.44	71.89	69.91	91.51	67.08	54.98
RGB + IR	76.97	73.05	89.85	69.21	53.04	76.78	76.77	**95.53**	82.54	58.00	71.41	69.86	90.37	67.00	55.56
IR only (Full)	77.13	**75.00**	**91.88**	69.84	**55.57**	**77.38**	**77.33**	95.45	**83.54**	**58.87**	**72.05**	**70.40**	**91.64**	**67.33**	**55.72**

Table 18: Sensitivity analysis of the bottleneck rank r of the IR refinement adapter. The best and second-best results in each column are highlighted in bold and underlined, respectively.

Adapter rank r	RefFLIR					RefM^{3}FD					RefMFAD				
	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C	Val	Test	Test-A	Test-B	Test-C
4	77.10	72.90	89.26	68.75	53.54	77.28	76.58	95.45	82.81	57.52	71.81	69.67	91.13	66.96	54.74
8	77.13	73.66	89.97	69.84	54.15	77.17	76.64	95.45	82.72	57.68	71.89	69.55	91.13	66.72	54.62
16	**77.13**	**75.00**	**91.88**	69.84	**55.57**	**77.38**	**77.33**	95.45	**83.54**	**58.87**	72.05	**70.40**	**91.64**	**67.33**	**55.72**
32	77.03	73.62	89.73	**70.09**	53.75	75.29	76.00	**95.69**	81.17	56.43	**72.23**	70.05	91.01	67.00	55.33

frequency-specific structural supervision to the IR branch, suggesting that teacher-guided IR refinement better compensates for the limited infrared modeling of RGB-centric VLMs and enables more effective use of complementary IR cues for grounding.

D.9 Sensitivity to the IR Refinement Adapter Rank

We examine the bottleneck rank r of the IR refinement adapter $\boldsymbol{f}_{\text{adapter}}$ in Eq. 5. Its down-projection maps features from d to r dimensions, and the up-projection restores the original dimension d. Table 18 compares $r \in \{4, 8, 16, 32\}$.

The validation results provide a consistent criterion for rank selection. Setting $r = 16$ achieves validation accuracies of 77.13%, 77.38%, and 72.05% on RefFLIR, RefM^{3}FD, and RefMFAD, respectively. It ties for the highest validation accuracy on RefFLIR, achieves the highest accuracy on RefM^{3}FD, and ranks second on RefMFAD. More importantly, it yields the highest mean validation accuracy of 75.52%, compared with 75.40%, 75.40%, and 74.85% for $r = 4$, 8, and 32, respectively. We therefore select $r = 16$ based solely on mean validation performance. Under this configuration, the model also achieves the highest Test and Test-C accuracy across all three datasets. Increasing the rank to 32 does not improve the mean validation accuracy or the primary Test and Test-C results, indicating that additional adapter capacity is unnecessary.

D.10 Model Complexity

Table 19 compares the parameter counts and computational costs of RGB-T grounding models and TCFDistill variants. All models are evaluated with paired 224×224 RGB–IR inputs, a batch size of one, and their respective text lengths. For MV-D-MDETR, $77 \rightarrow 40$ denotes the text lengths used by the encoder and fusion module, respectively. We report each model's total parameter count and the computational cost of its complete evaluation forward pass, including auxiliary outputs. For TCFDistill, the teacher model is excluded because it is used only during training.

Table 19: Parameter counts and computational costs of RGB-T grounding models and TCFDistill variants. Lower parameter counts and GFLOPs are preferred, with the lowest and second-lowest values highlighted in bold and underlined, respectively.

Method	Params (M) ↓	Text length	GFLOPs ↓
MV-D-MDETR	<u>163.40</u>	77 → 40	<u>81.32</u>
MV-MMCA	**153.952**	20	**25.73**
MV-HiVG-B	213.345	77	131.12
MV-OneRef-B	224.532	64	95.24
MV-HiVG-B + SARM	213.531	77	151.23
MV-HiVG-B + DFSD	213.724	77	141.10
TCFDistill ($r = 8$)	213.801	77	151.33
TCFDistill ($r = 16$)	213.809	77	151.34



Figure 6: Additional qualitative comparisons of RGB-T visual grounding in challenging scenes, including small targets, nighttime conditions, and rainy weather. Green and red boxes denote ground-truth and predicted bounding boxes, respectively.

With $r = 16$, TCFDistill contains 213.809M parameters and requires 151.336 GFLOPs. Relative to MV-HiVG-B, it adds 0.464M parameters (0.22%) and 20.214 GFLOPs (15.42%). The small parameter increase reflects the use of compact refinement and modulation modules. The additional computation mainly comes from processing the frequency tokens introduced by SARM and generating auxiliary predictions for both modalities in the current implementation.

The component variants separate these costs: adding SARM or DFSD individually increases computation by 20.105 and 9.974 GFLOPs, respectively, relative to MV-HiVG-B. Their combination adds 20.214 GFLOPs, with shared auxiliary computations counted once. Increasing the adapter rank from 8 to 16 adds approximately 0.003 GFLOPs, indicating that the rank has little effect on the overall computational cost.

Overall, TCFDistill offers a favorable trade-off between grounding accuracy and computational cost, improving localization performance with a 0.22% increase in parameters and a 15.42% increase in GFLOPs over MV-HiVG-B.

E Additional Qualitative Analysis

E.1 Qualitative Grounding under Challenging Conditions

Figure 6 compares TCFDistill with MV-OneRef, MV-TransVG, and RGBT-VGNet on selected RGB–IR scenes containing small, weakly visible, or partially occluded targets. These examples include low-light and rainy conditions in which RGB appearance cues are degraded. In the illustrated cases, TCFDistill produces bounding boxes that more closely match the ground truth. MV-OneRef and MV-TransVG exhibit missed targets or displaced predictions, while RGBT-VGNet recovers

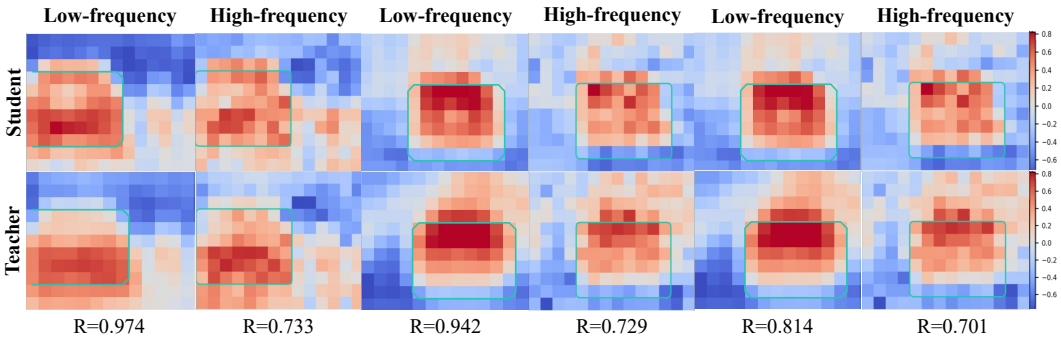


Figure 7: Qualitative visualization of the structural relations aligned by DFSD. LF-TAD aligns the low-frequency target-affinity maps a_s^l and a_t^l, while HF-SCD aligns high-frequency RGB–IR correspondence maps. Cyan boxes indicate ground-truth regions. Student and teacher maps share the same color scale.



Figure 8: Qualitative comparison of query-conditioned grounding heatmaps in challenging RGB–IR scenes. Solid boxes denote ground-truth regions, while dashed boxes denote predicted bounding boxes.

some targets with less precise localization. The localization advantage is particularly evident for targets occupying small image regions. Collectively, these examples highlight the benefit of combining DFSD for target-centric, frequency-specific structural transfer with SARM for consistency-guided modulation of shared structures and complementary responses.

E.2 Qualitative Analysis of DFSD

Figure 7 illustrates the structural relations transferred by DFSD. Here, R denotes the Pearson correlation coefficient between paired student and teacher relation maps, with higher values indicating greater spatial similarity. In LF-TAD, the student target-affinity map a_s^l exhibits a spatial organization similar to that of the teacher map a_t^l. Both maps assign higher affinity scores to regions corresponding to the referred target and lower scores to the surrounding background. The three examples yield $R = 0.974$, 0.942, and 0.814, respectively. This spatial similarity provides qualitative evidence that LF-TAD transfers query-conditioned low-frequency target-affinity relations.

In HF-SCD, the student and teacher high-frequency correspondence maps exhibit similar fine-grained spatial variations around the referred object, including responses near object boundaries and within target regions, with $R = 0.733$, 0.729, and 0.701 for the three examples. These observations complement the low-frequency results: LF-TAD captures target-level spatial organization, whereas HF-SCD captures fine-grained local RGB–IR correspondences. Together, the visualizations and

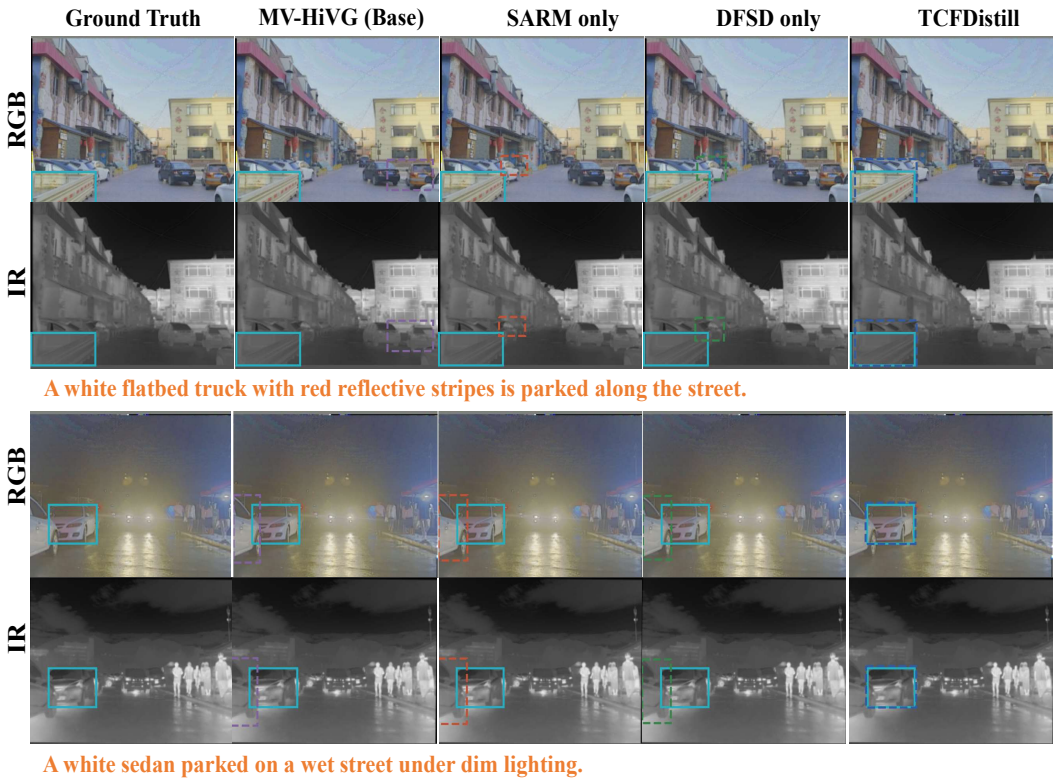


Figure 9: Qualitative comparison of grounding results under component ablations: Base, SARM only, DFSD only, and the full TCFDistill model. Ground-truth (GT) boxes are shown for reference.

correlation scores provide evidence that DFSD transfers complementary structural relations across low- and high-frequency components.

E.3 Additional Qualitative Comparison of Grounding Heatmaps

Figure 8 compares query-conditioned grounding heatmaps in challenging RGB–IR examples involving small targets, nighttime scenes, and low-visibility conditions. In the illustrated examples, TCFDistill produces spatial responses concentrated within or near the ground-truth target regions, with predicted boxes that more closely match the referred objects than those of the compared methods. RGBT-VGNet and MV-HiVG exhibit stronger responses to nearby background structures, such as road surfaces, illuminated areas, and surrounding vehicles, resulting in less accurate localization. These examples demonstrate that TCFDistill produces more target-aligned response distributions under challenging conditions. The improved target focus highlights the benefit of combining target-centric, frequency-specific structural transfer through DFSD with consistency-guided modulation of shared structures and complementary responses through SARM.

E.4 Qualitative Analysis of Component Complementarity

Figure 9 compares the grounding results of the Base model, the variants equipped with only SARM or DFSD, and the full TCFDistill model. Using either module alone can improve the localization over the Base model, but the predicted boxes may still deviate from the referred objects or cover irrelevant regions. When DFSD and SARM are used jointly, the predicted boxes are more accurately aligned with the target objects and better preserve their spatial extent. This complementarity is particularly evident under dim lighting, where weak RGB appearance cues make localization more challenging. In these cases, DFSD provides refined IR representations through target-centric structural transfer, while SARM further exploits local RGB–IR spectral interactions to improve the final grounding results.