T-CLIP: Enabling Thermal Perception for Contrastive Language-Image Pretraining

Anonymous authors
Paper under double-blind review

Abstract

Thermal imaging offers a powerful alternative to visible-spectrum vision under challenging conditions such as low illumination and adverse weather, yet foundational vision-language models like CLIP fail to align thermal images with textual descriptions due to a fundamental thermal perception gap. We identify three major challenges: the lack of captioned thermal datasets, the inability of standard LLMs to reason about thermal phenomena, and a key representational challenge in thermal imaging where global scene context and object-level heat signatures conflict when learned together in a single embedding space. To address these, we introduce IR-Cap, the first physics-aware thermal captioning pipeline and dataset providing complementary global and fine-grained thermal descriptions across three public benchmarks, and T-CLIP, a decoupled dual-LoRA framework that independently adapts CLIP for scene-level and object-level thermal understanding. T-CLIP achieves consistent improvements over all baselines across three thermal benchmarks in cross-modal retrieval, and we provide an exploratory demonstration of its applicability to text-conditioned thermal image generation.

1 Introduction

Deep learning has revolutionized computer vision, yet the field remains constrained by its heavy reliance on visible-spectrum data. This dependency on RGB imagery becomes a failure point in challenging conditions such as low-light environments, fog, smoke, or other visual obstructions. Thermal imaging, which captures infrared radiation emitted by objects, offers a powerful alternative (Vollmer & Möllmann, 2018). By visualizing heat signatures rather than reflected light, thermal cameras operate effectively in total darkness and penetrate various obscurants, making them indispensable for applications like 24/7 autonomous navigation, search-and-rescue operations, medical diagnostics, and surveillance systems (Farooq et al., 2023; Wilson et al., 2023; He et al., 2021; Gade & Moeslund, 2014; Nguyen et al., 2021; Hirsh et al., 2012). Despite these advantages, the full potential of thermal imaging remains largely untapped, as vision models lack the capability to interpret thermal data.

This limitation is particularly evident in Vision-Language Models (VLMs) such as CLIP Radford et al. (2021). We identify a significant *thermal perception gap* in CLIP's understanding, likely stemming from its exclusive training on RGB image-text pairs. Unlike conventional visible-spectrum imagery, where appearance is determined by color and texture, thermal representations are governed by physical properties such as temperature, emissivity, and radiative balance that are largely absent in CLIP's pretraining distribution. Our analysis reveals that this discrepancy extends beyond a superficial domain shift, manifesting as specific representational limitations. For instance, when guiding generative models like SDXL Rombach et al. (2022); Podell et al. (2023b) with prompts such as *"an infrared thermal image of a road with lane markings,"*, the outputs often resemble pseudo-colored RGB images, lacking the characteristic thermal signatures expected in infrared images. This occurs because the CLIP text encoder relies on optical descriptors like "white" and "yellow" for lane markings, failing to account for the fact that these features are visible in thermal images due to material emissivity rather than reflected light. Similarly, while CLIP vision encoder can recognize objects via shape contours, it is often insensitive to the underlying heat signatures, which is the primary

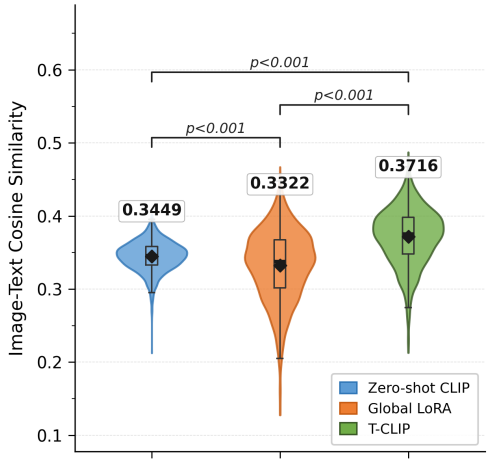

Image-text cosine similarity between matched thermal image-caption pairs measures how well each model aligns a thermal image with its correct description in embedding space. Zero-shot CLIP (cosine similarity = 0.3449) shows insufficient thermal image-text alignment, resulting in negligible retrieval performance (R@1 = 0.003). Standard LoRA fine-tuning of CLIP on thermal data (Global LoRA) shows a marginal drop in cosine similarity (0.3449 → 0.3322) compared to zero-shot CLIP, reflecting instability when adapting a model trained on RGB data to the thermal domain, using only generic scene descriptions without explicit thermal supervision. T-CLIP, through decoupled dual-context alignment on physics-aware captions, achieves the strongest alignment (0.3716), corresponding to a 26× improvement in R@1 (0.003 → 0.078; Table 1). All pairwise differences are statistically significant ($p \ll 0.001$, Welch t-test, $n = 2252$); p-values are unchanged after Benjamini-Hochberg correction (FDR = 0.05), confirming robustness to multiple comparison adjustment (appendix section A.1).

Figure 1: Mean image-text cosine similarity of matched thermal image-caption pairs on the KAIST test set (Hwang et al., 2015) ($n = 2252$), reflecting the thermal perception gap in standard CLIP (Radford et al., 2021) and T-CLIP's improvement.

discriminative signal in the thermal domain. Consequently, the model struggles to differentiate between physically different thermal states, such as a "warm, running engine" versus a "cool, parked car."

We quantify this thermal perception gap by measuring the mean image-text cosine similarity between matched thermal image-caption pairs. As shown in Figure 1, zero-shot CLIP achieves a cosine similarity of merely 0.3449 on the KAIST (Hwang et al., 2015) test set (2252 image-text pairs), confirming that thermal images and their descriptions are poorly aligned in CLIP's embedding space. This directly translates to negligible retrieval performance (R@1 = 0.003, Table 1), indicating that CLIP's limitations on thermal data go beyond a simple domain shift, pointing to a deeper cross-modal representation gap in the thermal domain. All pairwise cosine similarity differences are statistically significant ($p \ll 0.001$, Welch t-test, $n = 2252$, BH-corrected; appendix section A.1).

Bridging this thermal perception gap requires addressing three key challenges. First, the thermal vision community lacks large-scale captioned datasets; existing datasets are primarily designed for detection or segmentation and lack descriptive textual annotations. Even the IR-500 (Ran et al., 2025) dataset assigns a generic caption—*"An infrared thermal image"*, to all samples, omitting the rich thermal physics underlying the imagery. Second, standard LLMs fail to generate thermally informative descriptions as they default to RGB-style priors. This bias causes them to hallucinate color and texture attributes while overlooking essential thermal features such as heat signatures, material emissivity, and temperature relationships. Third, our experiments reveal an important insight into the nature of thermal representations. Thermal images carry two distinct levels of information—global-level scene content and object-level heat signatures, that cannot be effectively mapped within a single CLIP embedding space. We demonstrate that simultaneous optimization leads to representational interference, while sequential adaptation results in catastrophic forgetting (Tables 4 and 5).

To address these challenges, we propose T-CLIP, a simple yet effective framework that bridges the thermal perception gap through two primary technical contributions. First, we develop IR-Cap, a physics-aware thermal captioning pipeline that leverages paired RGB images as semantic anchors, enabling Qwen2.5-VL-72B-Instruct (Wu et al., 2025) to reason about the underlying thermal properties of a scene that cannot be directly inferred from thermal images by standard LLMs. Using specialized dual prompts, we guide the model to generate two complementary caption types: one capturing the global scene context and the other focusing on object-level thermal attributes such as material emissivity, heat retention, and active heat signatures. This

approach circumvents the inherent inability of standard LLMs to interpret thermal phenomena, producing descriptions that are both semantically rich and thermally relevant. Second, to address the representational challenges specific to thermal data, we propose a decoupled dual LoRA framework that independently trains two specialized branches, one dedicated to scene-level thermal context and another to object-level heat signature understanding, on their respective caption types, and fuses their complementary representations at inference via a weighted ensemble. This decoupled design consistently outperforms single-branch variants across retrieval metrics and datasets, demonstrating that scene-level and object-level thermal representations capture complementary information that benefits from specialized rather than joint optimization.

Our contributions can be summarized as follows:

- We introduce **IR-Cap**, the first thermal caption dataset providing physics-aware descriptions beyond generic modality labels. Using paired RGB images as semantic anchors, we develop a dual-prompt captioning pipeline that generates two complementary caption types: Global Thermal Captions capturing scene-level context, and Fine-Grained Thermal Captions encoding object-level heat signatures. The pipeline applies to any paired RGB-thermal dataset; in this work we annotate KAIST, FLIR, and FMB, releasing both annotations and pipeline publicly.
- We present **T-CLIP**, a decoupled dual-LoRA framework for thermal CLIP adaptation. Through controlled ablations and feature space analysis, we show that scene-level thermal context and object-level heat signatures are geometrically divergent and interfere under joint optimization, motivating a decoupled design that leverages IR-Cap's complementary dual captions.
- T-CLIP achieves consistent improvements over all baselines across three thermal benchmarks in image-to-text and text-to-image retrieval. Beyond retrieval, we provide an exploratory demonstration showing that the thermally adapted text encoder can be integrated into SDXL for text-conditioned thermal image generation, motivating future systematic study.

2 Related Works

CLIP. Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) has established itself as a foundational Vision-Language Model, learning a unified embedding space from extensive RGB image-text pairs. Its zero-shot capabilities have enabled widespread adoption in detection (Gu et al., 2022; Li et al., 2021), segmentation (Xu et al., 2022; Li et al., 2022a), and generative modeling (Ramesh et al., 2022; Frans et al., 2022). To avoid the cost of full retraining, lightweight adaptation methods have emerged: CoOp (Yao et al., 2023) optimizes soft prompt tokens for classification, while CLIP-Adapter (Gao et al., 2024) introduces bottleneck layers for efficient domain transfer. Despite these advances, CLIP's representations remain intrinsically tied to RGB-based visual semantics, lacking the physical grounding required for non-visible modalities. While specialized adaptations exist for domains such as medical imaging (Lai et al., 2023), the specific challenge of bridging the thermal perception gap remains unaddressed, leaving foundational VLMs insensitive to the underlying physics of thermal images.

Vision-Language Datasets. The progress in multimodal learning has been fueled by large-scale, captioned datasets in the visible spectrum, such as COCO Lin et al. (2014), Visual Genome Krishna et al. (2017), and LAION Schuhmann et al. (2022). In stark contrast, the thermal imaging community suffers from a severe scarcity of analogous resources. Existing thermal datasets are primarily designed for tasks like classification and object detection, offering only categorical labels (e.g., "person," "car") (Hwang et al., 2015; FLIR Systems, 2025; Liu et al., 2023). Datasets that do include text often provide only generic, modality-focused descriptions, such as "an infrared image" (e.g., IR-500 (Ran et al., 2025)), which fail to capture the rich, physics-based attributes like heat signatures and material emissivity. This absence of a large-scale, physics-aware thermal caption dataset is a primary bottleneck for developing thermal VLMs, which IR-Cap addresses.

Thermal Image Generation. Most literature on thermal image synthesis formulates the task as an image-to-image translation problem, converting RGB frames to infrared equivalents (Iwashita et al., 2023;

Kniaz et al., 2017; Qazi et al., 2025). Text-conditioned generation remains largely unexplored; while DiffV2IR (Ran et al., 2025) provides a notable exception, it relies on generic captions and auxiliary spatial conditioning (RGB and segmentation maps) at inference. This necessitates intensive end-to-end training and limits the model's ability to reason about thermal physics from text alone. In contrast, T-CLIP offers a proof-of-concept for a thermal generation approach only from text prompts without any additional spatial priors.

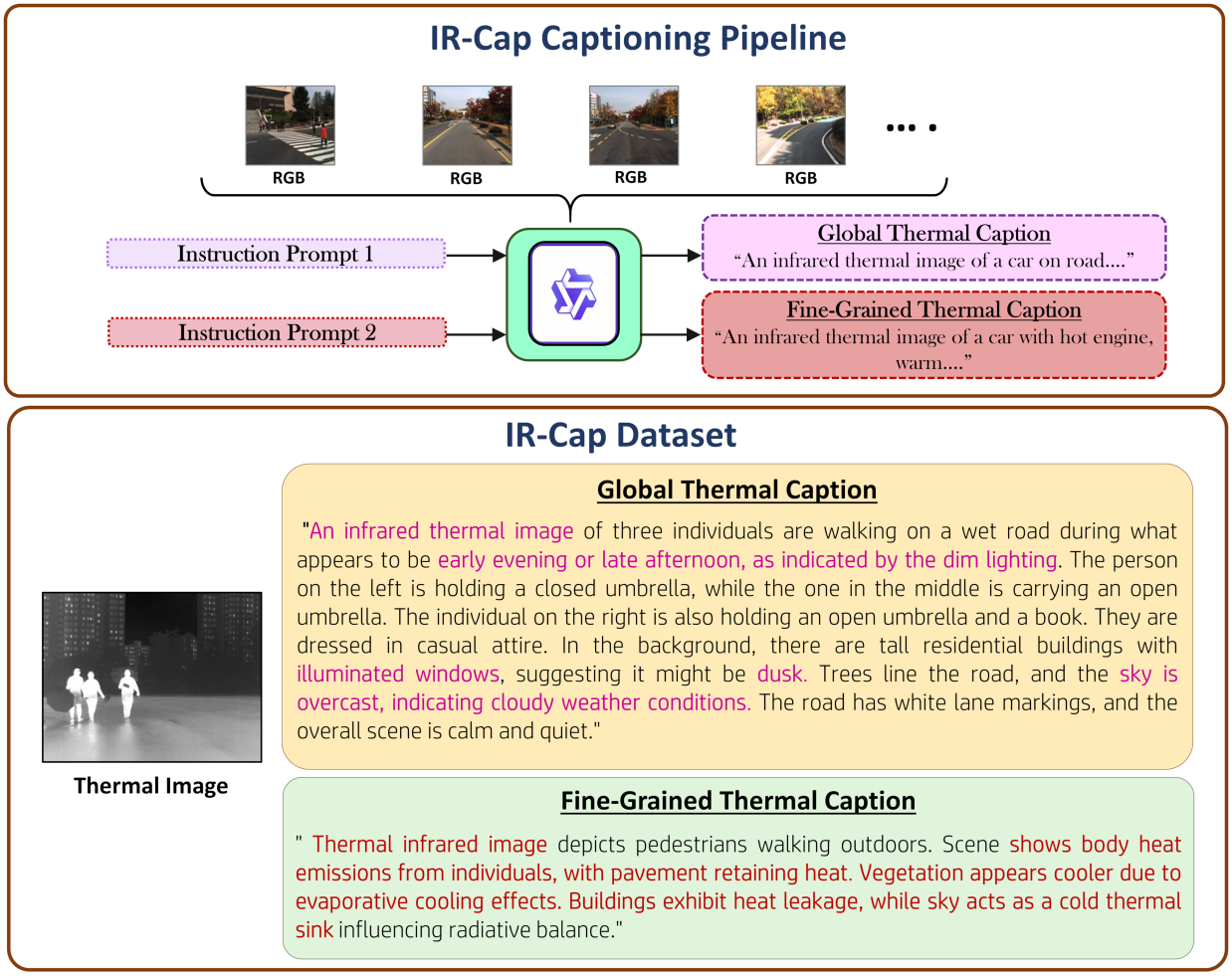

Figure 2: IR-Cap captioning pipeline and dataset. (Top): The IR-Cap captioning pipeline leverages paired RGB images as semantic anchors to generate complementary physics-aware captions for thermal images. Since standard vision-language models are trained on RGB data and exhibit limited understanding of infrared thermal characteristics, we condition Qwen2.5-VL-72B-Instruct on the RGB counterpart of each thermal image and use two specialized instruction prompts to steer the model toward reasoning about thermal attributes. Instruction Prompt 1 elicits a global thermal caption capturing scene-level environmental context, while Instruction Prompt 2 elicits a fine-grained thermal caption encoding object-level heat signatures and thermal phenomena. (Bottom): A representative IR-Cap dataset sample showing a thermal image alongside its two complementary physics-aware captions generated by the pipeline.

3 IR-Cap Dataset

While conventional RGB images can be described through direct visual attributes such as color and texture, thermal appearances are governed by physical properties such as temperature, emissivity, and environmental

context, which are not directly observable in the visible spectrum. This semantic-physical gap has limited the development of effective thermal-visual reasoning systems. To address this, we introduce IR-Cap (Figure 2), the first thermal caption dataset providing physics-aware descriptions paired with thermal images across three public benchmarks.

Motivation and Captioning Strategy. Since standard LLMs cannot reliably interpret infrared imagery, we exploit the paired visible-spectrum images available in existing multispectral datasets as semantic context. Rather than presenting the thermal image directly to the model, we condition Qwen2.5-VL-72B-Instruct (Wu et al., 2025) on its RGB counterpart, allowing it to anchor scene semantics, like object identities, spatial layout, and environmental context, while our specialized dual prompts redirect its reasoning toward the thermal implications of the observed scene, rather than its visible-spectrum appearance.

We interpret thermal imagery as comprising two complementary semantic levels: (i) global scene-level understanding and (ii) localized object-centric thermal signatures. We observe that scene-level attributes, including environmental context, weather conditions, and coarse activity semantics, can often be inferred from aligned RGB imagery and described reliably by modern vision-language models trained predominantly on visible-spectrum data. In contrast, fine-grained thermal phenomena, such as heat intensity, emissivity variations, engine activity, and relative temperature distributions, are not directly observable from RGB appearance alone. Consequently, standard vision-language models lack the intrinsic thermal priors required to accurately characterize such infrared-specific properties. To address this limitation, we design specialized prompts that encourage the model to reason about plausible thermal behavior at the object level, enabling the generation of thermally descriptive annotations, such as distinguishing between the hot engine region of a moving vehicle and the comparatively cooler engine of a parked vehicle. These complementary descriptions collectively serve as physics-aware semantic priors that bridge the gap between visible-spectrum vision-language models and the thermal domain. This perspective inspires our dual prompting strategy, described below.

Dual Prompting Strategy. We design two distinct instruction prompts, each targeting a different level of thermal understanding:

Instruction Prompt 1 (Global Thermal Captions, c_g): This prompt instructs the model to describe the scene from a macro-level thermal perspective, focusing on environmental factors that govern the overall heat distribution. Specifically, it asks the model to reason about illumination conditions, time of day, weather patterns, and material composition comprising of factors that collectively determine the thermal landscape of the scene. For example, a night scene with wet roads would prompt descriptions of reduced ambient heat and increased thermal contrast between warm objects and cool surfaces.

Instruction Prompt 2 (Fine-Grained Thermal Captions, c_{fg}): This prompt directs the model toward object-level thermal reasoning, focusing on heat signatures, material emissivity differences, and temperature relationships between specific objects. For example, for the same night scene, this prompt elicits descriptions of individual objects' thermal behavior. For example, "moving vehicles exhibit hot engines and warm tires, while pedestrians emit body heat" , capturing the discriminative thermal attributes not present in the global descriptions.

Both prompts include the keywords "infrared" and "thermal" in the caption prefix to provide a modality-specific tag, ensuring the generated descriptions are framed within the thermal domain rather than defaulting to RGB-style language. The full prompt templates are provided in the appendix section A.2.

Dataset Statistics Using this pipeline, we annotate three public thermal benchmarks, KAIST (Hwang et al., 2015), FLIR (FLIR Systems, 2025), and FMB (Liu et al., 2023), augmenting each with paired (c_g, c_{fg}) physics-aware captions. To quantify attribute coverage, we perform keyword-based analysis over the full caption corpus, checking for the presence of predefined attribute categories (illumination, time of day, weather, materials for Global captions; heat signatures and thermal phenomena for Fine-Grained captions) and reporting the fraction of captions that mention each attribute. As illustrated in Figure 3, this analysis confirms the broad and complementary coverage of our dual-prompting strategy. *Global Thermal Captions* provide necessary context regarding ambient environmental conditions, with near-universal coverage of illumination

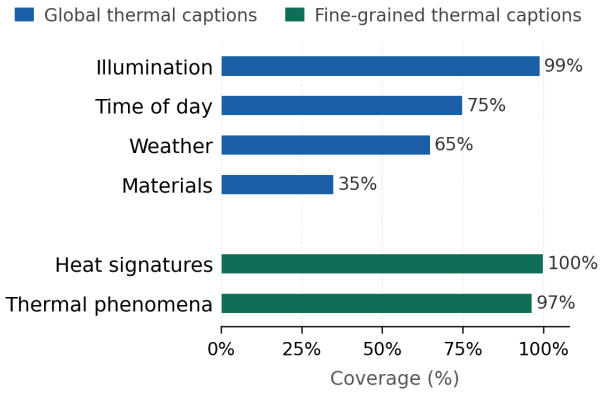

Figure 3: Attribute coverage of IR-Cap captions. Values indicate the percentage of captions within each type that mention the corresponding attribute.

(99%) and significant inclusion of temporal and meteorological context (time of day: 75%, weather: 65%). In contrast, *Fine-Grained Thermal Captions* capture intrinsic thermal properties exclusively: 100% of these captions identify heat signatures, while 97% articulate complex phenomena such as emissivity variations and radiative balance. Annotating thermal images with such physics-aware metadata provides rich thermal priors that enable comprehensive scene-level contextual understanding and fine-grained object-level reasoning, thereby serving as an essential supervisory signal for mitigating the thermal perception gap identified in section 4.1.

4 Method

We begin in section 4.1 by characterizing two fundamental challenges in thermal vision-language alignment: the thermal perception gap in CLIP (Figure 1), and the empirically revealed geometric divergence between global and fine-grained thermal feature spaces (Figure 4), which together motivate our design. Section 4.2 formalizes the problem, after which section 4.3 presents our decoupled dual-LoRA training strategy, illustrated in Figure 5. Finally, Figure 4.4 describes the weighted feature fusion at inference (Figure 6).

4.1 Challenges in Thermal Vision-Language Alignment

Thermal Perception Gap. Standard CLIP, pre-trained on RGB-centric web data, exhibits a significant semantic gap when applied to thermal imagery. As shown in Figure 1, zero-shot CLIP achieves a mean image-text cosine similarity of only 0.3449 on the KAIST (Hwang et al., 2015) thermal dataset, despite being paired with carefully curated thermal captions. This is not merely a distributional shift; thermal images encode physically unique signals (emissivity, heat flux, material conductivity) that are entirely absent from CLIP's training distribution. Strikingly, single-LoRA adaptation on global captions (*Global LoRA*, mean 0.3322) degrades alignment relative to zero-shot CLIP, and all pairwise differences are statistically significant ($p \ll 0.001$, Welch t-test, BH-corrected; appendix section A.1), confirming that naïve fine-tuning strategies are insufficient. Effective thermal alignment requires a purpose-built representation strategy.

Complementary yet Divergent Thermal Semantics. Thermal imagery encodes information at two conceptually distinct semantic levels: *global scene characteristics*, encompassing environmental conditions, time of day, spatial layout, and material composition, which partially overlap with visible-spectrum cues, and *object-level thermal physics*, including emissivity variations, localized heat signatures, and thermal conductivity differences, which are exclusively observable in the thermal domain and demand physics-aware representation. We empirically reveal these two levels to be not only semantically complementary but geometrically divergent in feature space, by training two LoRA branches independently — one exclusively on global scene captions and the other exclusively on fine-grained thermal captions and probing their respective

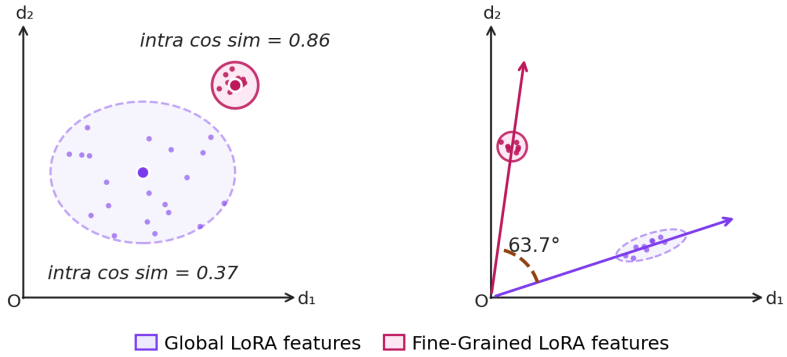

Figure 4: Feature geometry of independently trained global and fine-grained LoRA branches on KAIST (Hwang et al., 2015). (Left) Global LoRA features exhibit low intra-class cohesion (intra cosine similarity = 0.37), while fine-grained LoRA features form tight compact clusters (= 0.86). (Right) Mean feature vectors subtend 63.7°, confirming geometric divergence between the two representation spaces and motivating the decoupled dual-LoRA design (see section 4.1).

feature distributions as visualized in Figure 4, global LoRA features exhibit low intra-class cohesion (mean intra cosine similarity = 0.37), reflecting the inherent diversity of scene-level thermal contexts, while fine-grained LoRA features form tight, compact clusters (mean intra cosine similarity = 0.86), consistent with the physically constrained nature of object-level heat signatures. Furthermore, the mean feature vectors of the two branches subtend an angle of 63.7°, confirming that global and fine-grained thermal representations occupy significantly different regions of the embedding space. Joint optimization of a single LoRA on mixed captions conflates these divergent objectives, producing gradient interference that degrades both, as shown in Table 4. Sequential training partially mitigates interference but leaves the second-stage LoRA without awareness of the first-stage objectives, leading to suboptimal joint representations, as confirmed by Table 5. These findings motivate a fully decoupled strategy with independent parameter spaces, which we describe next.

4.2 Problem Formulation

Given a thermal image $I_{\text{th}} \in \mathbb{R}^{H \times W \times 1}$ and two paired captions — a global thermal caption c_g describing environmental context and a fine-grained caption c_{fg} encoding object-specific heat signatures, we aim to learn, over paired-caption dataset $\mathcal{D}$, complementary representation spaces that capture both semantic levels without interference. We introduce two independent Low-Rank Adapter (LoRA) modules θ_g and θ_f, each applied to both the vision and text encoders of a frozen CLIP (Radford et al., 2021) backbone, and optimize them separately:

$$\theta_g^* = \arg\min_{\theta_g} \; \mathbb{E}_{(I_{\text{th}}, c_g) \sim \mathcal{D}} \left[\mathcal{L}_{\text{global}} \big(F_{\theta_g}(I_{\text{th}}),\, G_{\theta_g}(c_g) \big) \right] \tag{1}$$

$$\theta_f^* = \arg\min_{\theta_f} \; \mathbb{E}_{(I_{\text{th}}, c_{fg}) \sim \mathcal{D}} \left[\mathcal{L}_{\text{fg}} \big(F_{\theta_f}(I_{\text{th}}),\, G_{\theta_f}(c_{fg}) \big) \right] \tag{2}$$

where $F_{\theta_g}(\cdot) = F_{\text{frozen}}(\cdot) + \Delta F_{\theta_g^v}(\cdot)$ and $G_{\theta_g}(\cdot) = G_{\text{frozen}}(\cdot) + \Delta G_{\theta_g^t}(\cdot)$ denote the base encoders augmented with their respective LoRA adaptations, and analogously for θ_f. Since the two modules are optimized independently over disjoint supervisory signals, with the shared CLIP backbone kept frozen throughout, there is zero gradient interference between the two representation spaces.

Beyond the geometric motivation established in section 4.1, LoRA adaptation is particularly suited to the thermal domain, where large-scale paired datasets are scarce; full fine-tuning of CLIP would therefore risk catastrophic forgetting of its rich visual-semantic priors given the limited thermal training data, whereas

LoRA confines adaptation to a small set of low-rank parameters while preserving the frozen backbone's generalizable representations.

To our knowledge, T-CLIP is the first framework to explicitly decompose thermal image understanding into global scene context and object-level heat signature physics, empirically reveal their geometric divergence in feature space (Figure 4), and leverage this finding as an architectural prior instantiated through independent LoRA parameter spaces on a shared frozen CLIP backbone, with ablations confirming the necessity of this decoupling (Tables 4 and 5).

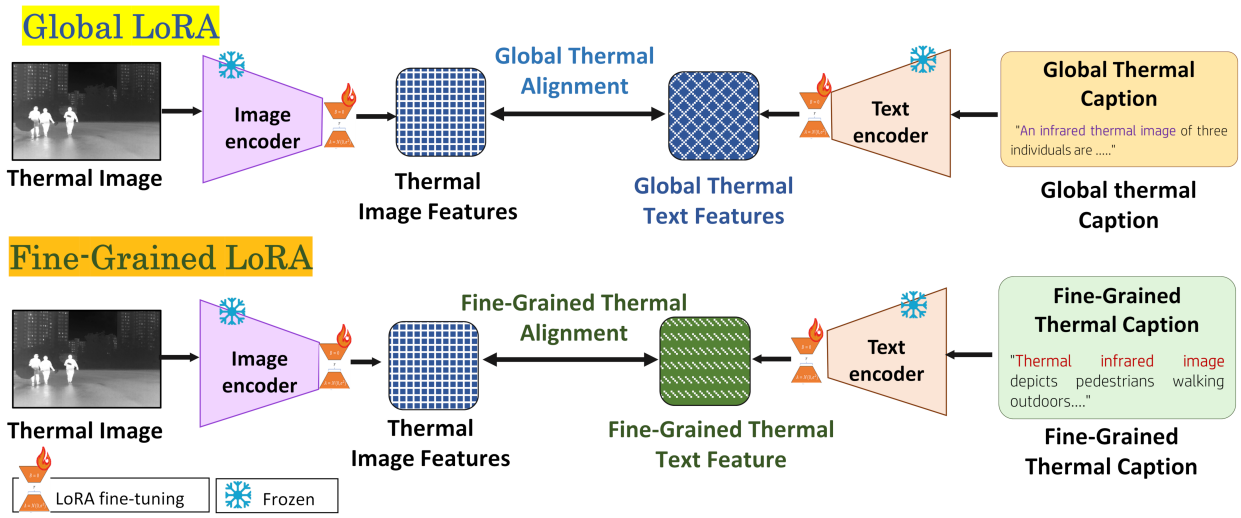

Figure 5: T-CLIP dual LoRA training pipeline. Two independent LoRA modules on a frozen CLIP backbone are optimized on semantically distinct caption types, preventing gradient interference while enabling complementary thermal representation learning.

4.3 Dual LoRA Training for Decoupled Thermal Representation

The overall training pipeline is illustrated in Figure 5, with the corresponding pseudocode provided in section A.6.3 of the appendix. Each LoRA branch is trained independently with a standard InfoNCE contrastive objective (Radford et al., 2021), operating on the same set of thermal images but with semantically distinct caption types.

Global Thermal LoRA. θ_g is trained exclusively on global captions c_g describing scene content, illumination, weather, material composition, and temporal context:

$$\mathcal{L}_{\text{global}}(\theta_g) = -\log \frac{\exp\big(\text{sim}(F_{\theta_g}(I_{\text{th}}),\, G_{\theta_g}(c_g))\,/\,\tau\big)}{\sum_{k=1}^{B} \exp\big(\text{sim}(F_{\theta_g}(I_{\text{th}}),\, G_{\theta_g}(c_g^k))\,/\,\tau\big)} \tag{3}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a learned temperature parameter, B is the batch size, and c_g^k denotes the k-th global caption in the batch serving as a negative sample. The same notation applies to c_{fg}^k, (the k-th fine-grained caption) in $\mathcal{L}_{\text{fg}}$ (Equation 4). This enforces θ_g to encode the macroscopic thermal landscape — spatial heat flow, environmental gradients, and scene-level structure, without contamination from object-level supervisory signal.

Fine-Grained Thermal LoRA. θ_f is trained exclusively on c_{fg}, captions describing localized thermal phenomena such as emissivity contrasts, heat source attribution, and object-specific temperature anomalies

(*e.g.*, "hot engine block," "warm pedestrian torso"):

$$\mathcal{L}_{\text{fg}}(\theta_f) = -\log \frac{\exp\big(\text{sim}(F_{\theta_f}(I_{\text{th}}),\, G_{\theta_f}(c_{fg}))\,/\,\tau\big)}{\sum_{k=1}^{B} \exp\Big(\text{sim}(F_{\theta_f}(I_{\text{th}}),\, G_{\theta_f}(c_{fg}^{k}))\,/\,\tau\Big)} \tag{4}$$

The decoupled weight spaces preserve the geometric divergence observed in Figure 4 rather than collapsing it, and gradients from neither branch interfere with the other throughout training.

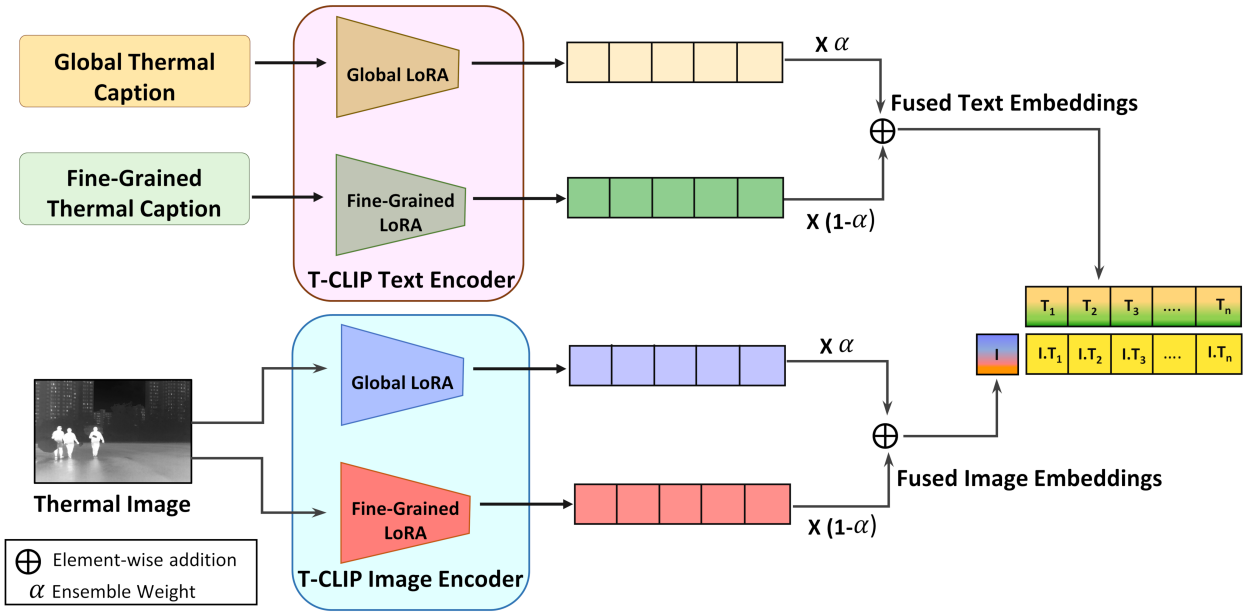

Figure 6: T-CLIP inference. Features from the global and fine-grained LoRA branches are combined via a scalar hyperparameter α, enabling flexible control over scene-level versus object-level emphasis at retrieval time.

4.4 Inference via Weighted Feature Fusion

Figure 6 illustrates the inference pipeline. At test time, both LoRA branches independently process a thermal image and their outputs are fused via a scalar hyperparameter α:

$$v_{\text{fused}} = \alpha \cdot F_{\theta_g}(I_{\text{th}}) + (1-\alpha) \cdot F_{\theta_f}(I_{\text{th}}) \tag{5}$$

For text-side fusion during text-to-image retrieval, captions from both branches are combined symmetrically:

$$w_{\text{fused}} = \alpha \cdot G_{\theta_g}(c_g) + (1-\alpha) \cdot G_{\theta_f}(c_{fg}) \tag{6}$$

$\alpha \in [0, 1]$ controls the relative contribution of each branch; we set $\alpha = 0.8$ by default, a value determined empirically (Figure 7), reflecting that global thermal context provides the dominant discriminative signal while fine-grained features supply complementary specificity. This fusion requires only a weighted sum of two forward passes, adding negligible computational overhead at inference relative to standard CLIP (Radford et al., 2021).

5 Experiments

We evaluate T-CLIP on thermal image-text retrieval (section 5.2) as our primary task and further explore its thermal image generation capability (section 5.7).

5.1 Experimental Settings

Evaluation Datasets. We evaluate on three publically available thermal imaging benchmarks. **KAIST** (Hwang et al., 2015) is a large-scale multispectral pedestrian dataset from which we use 50,000 unique thermal images spanning diverse outdoor and urban driving scenarios across both day and night conditions. **FLIR** (FLIR Systems, 2025) is a driving-focused thermal dataset; following (Zhang et al., 2020), we use the aligned version with 4,128 training and 1,013 validation pairs. **FMB** (Liu et al., 2023) is a multimodal benchmark augmented with 12,200 thermal images following the procedure of (Mayr et al., 2024). All results are reported on standard test splits to ensure fair comparison across methods.

For each dataset, global thermal captions and fine-grained thermal captions were generated using our proposed IR-Cap captioning pipeline (described in section 3), which produces paired global scene-level and object-level thermal descriptions for every image.

Evaluation Settings. We report Recall@K (R@K) for $K \in \{1, 5, 10, 25, 50, 100\}$, measuring the percentage of queries for which the correct match appears within the top-K retrieved results. We evaluate in both image-to-text (I2T) and text-to-image (T2I) directions across all three datasets. All evaluation protocols are held constant across methods and datasets. For each method and dataset, we report the mean and standard deviation of R@K over three independent training runs with different random seeds (0, 42, 123).

Training Settings. T-CLIP is fine-tuned for 10,000 steps with a batch size of 64 on a single NVIDIA A6000 GPU, requiring approximately 1.5 hours of training, highlighting the practical viability of our lightweight LoRA-based approach. The two LoRA branches (θ_g and θ_f) are trained independently on their respective caption types, with all hyperparameters held constant across both training runs. Detailed configurations, including learning rate, LoRA rank, and optimizer settings, are listed in appendix section A.6.

5.2 Comparing with Baselines

We compare T-CLIP against the following baselines, all evaluated on captions generated by our IR-Cap pipeline to ensure a fair and controlled comparison.

Zero-shot CLIP. We evaluate the off-the-shelf CLIP (Radford et al., 2021) model without any thermal adaptation, serving as a domain-agnostic lower bound. Since IR-Cap generates two caption types, we report two zero-shot CLIP variants: *CLIP (Global)*, evaluated on global scene captions, and *CLIP (F-G)*, evaluated on fine-grained thermal captions.

CLIP-Adapter. We include two CLIP-Adapter (Gao et al., 2024) variants: *CLIP-Adapter (Global)*, trained and evaluated on global thermal captions, and *CLIP-Adapter (F-G)*, trained and evaluated on fine-grained thermal captions. Both learn a lightweight residual feature adapter on top of frozen CLIP features without modifying the encoder weights, serving as an alternative parameter-efficient adaptation baseline.

DeCLIP. Full fine-tuning of CLIP on small thermal datasets risks catastrophic forgetting of pretrained visual-semantic representations. As a representative full-parameter baseline, we therefore include DeCLIP (Li et al., 2022b), a data-efficient extension that trains all CLIP parameters using image–image and text–text nearest-neighbour consistency via a momentum encoder. We report two variants: *DeCLIP (Global)*, trained on global thermal captions, and *DeCLIP (F-G)*, trained on fine-grained thermal captions.

Single LoRA Baselines. We fine-tune two independent single-branch LoRA models on the IR-Cap dataset: *Global LoRA*, fine-tuned and evaluated exclusively on global scene captions, and *Fine-Grained LoRA*, fine-tuned and evaluated exclusively on fine-grained thermal captions. These isolate the contribution of each caption type and directly motivate our dual-branch design.

T-CLIP Variants. We report three inference variants of T-CLIP, all using the fused dual-LoRA representation with $\alpha = 0.8$. In *T-CLIP (Global)*, a single global caption is passed through both LoRA branches

at inference, with the resulting branch features fused to produce the final representation. *T-CLIP (F-G)* follows the same protocol, substituting the global caption with a fine-grained thermal caption; and in *T-CLIP (Dual)*, our proposed model, the Global LoRA branch independently processes the global caption while the Fine-Grained LoRA branch processes the fine-grained caption, with both fused to obtain the final embedding. In scenarios where only a single caption type is available, *T-CLIP (Global)* and *T-CLIP (F-G)* serve as natural fallbacks; the available caption is routed through both branches without any architectural modification, preserving the dual-branch fusion mechanism and ensuring applicability across standard image-text retrieval settings where paired captions are not provided a priori.

Results are reported in Tables 1, 2, and 3; detailed mean $\pm$ std across all recall thresholds are provided in section A.4 of the appendix.

Across all three datasets, zero-shot CLIP (Radford et al., 2021) yields consistently poor retrieval performance, confirming that standard RGB-trained representations are fundamentally inadequate for thermal image-text alignment irrespective of whether global or fine-grained captions are used as queries.

Parameter-efficient adaptation substantially improves over zero-shot CLIP, with CLIP-Adapter (Gao et al., 2024), DeCLIP (Li et al., 2022b), and single-branch LoRA baselines all showing clear gains. DeCLIP achieves intermediate performance, generally surpassing CLIP-Adapter but falling short of LoRA-based methods across datasets. Global LoRA consistently and significantly outperforms both CLIP-Adapter and DeCLIP across all metrics and datasets, indicating that for thermal domain alignment, encoder weight modification via LoRA is more effective than residual feature adaptation or data-efficient self-supervision. Among single-branch methods, Global LoRA outperforms Fine-Grained LoRA across all datasets, suggesting that scene-level context provides a stronger retrieval signal in isolation, consistent with the feature geometry analysis in Figure 4.

T-CLIP (Dual) outperforms all baselines across the majority of metrics and datasets, demonstrating that the complementary thermal semantics captured by the two decoupled LoRA branches are mutually beneficial at retrieval time. On KAIST (Table 1), T-CLIP (Dual) achieves relative R@1 gains of +10.5% (I2T) and +21.9% (T2I) over Global LoRA, the strongest single-branch baseline, with consistent improvements on FLIR (Table 2) (+17.1% I2T, +10.4% T2I). Importantly, comparing T-CLIP (Global) and T-CLIP (F-G) against their single-branch LoRA counterparts reveals that the dual-LoRA representation enriches the visual embedding beyond what either caption type achieves independently. This confirms that training on complementary thermal semantics improves the quality of the learned representation space itself, not merely the inference-time fusion.

On FMB (Table 3), T-CLIP (Global) marginally outperforms T-CLIP (Dual) at lower I2T recall thresholds (R@1–R@10), though both variants consistently surpass Global LoRA baseline in this range. T-CLIP (Dual) recovers the lead at R@25 and above for I2T and dominates across all T2I metrics, consistent with the trend observed on KAIST and FLIR.

5.3 Ablation Study

Effect of Caption Type. Table 4 ablates the effect of caption type on retrieval performance on KAIST. Global LoRA substantially outperforms Fine-Grained LoRA, confirming that scene-level thermal context is more discriminative than object-level heat signatures in isolation, consistent with the feature geometry analysis in Figure 4. Training a single LoRA on an equal mixture of global and fine-grained captions (Mixed 50:50) reveals that it degrades performance on global captions relative to specialized Global LoRA, yet surpasses specialized F-G LoRA when evaluated on fine-grained captions. This suggests that exposure to global captions during training enriches the representation for fine-grained retrieval, but the inclusion of fine-grained captions simultaneously dilutes the global discriminability. This trade-off demonstrates that the two caption types impose conflicting supervision signals under shared training, motivating our decoupled design in T-CLIP where both caption types are optimized independently rather than under joint supervision.

Single vs. Cross-Initialized vs. Dual LoRA. Table 5 compares single-branch LoRA, cross-initialized LoRA variants, and our decoupled Dual LoRA across different caption evaluation settings. Dual LoRA with

Table 1: Text-image retrieval recall@K on the KAIST (Hwang et al., 2015) test set. R@1 reported as mean ± std across three independent training runs; all other R@K are means. **Bold**: best result. Underline: second best. F-G = fine-grained. Δ: relative R@K gain of T-CLIP (Dual) over the strongest baseline (Global LoRA).

	IMAGE-TO-TEXT RETRIEVAL						TEXT-TO-IMAGE RETRIEVAL					
METHOD	**R@1**	**R@5**	**R@10**	**R@25**	**R@50**	**R@100**	**R@1**	**R@5**	**R@10**	**R@25**	**R@50**	**R@100**
CLIP (Global)	0.003	0.016	0.034	0.068	0.110	0.179	0.002	0.013	0.027	0.048	0.080	0.135
CLIP (F-G)	0.001	0.004	0.008	0.029	0.057	0.096	0.001	0.007	0.014	0.030	0.048	0.088
CLIP-Adapter (Global)	0.018±0.001	0.072	0.123	0.229	0.357	0.521	0.020±0.002	0.070	0.119	0.234	0.360	0.538
CLIP-Adapter (F-G)	0.009±0.002	0.041	0.081	0.167	0.283	0.436	0.010±0.000	0.046	0.082	0.166	0.281	0.445
DeCLIP (Global)	0.038±0.002	0.129	0.206	0.349	0.497	0.644	0.041±0.000	0.133	0.209	0.357	0.496	0.650
DeCLIP (F-G)	0.013±0.000	0.048	0.088	0.169	0.273	0.433	0.014±0.000	0.056	0.094	0.175	0.285	0.439
Global LoRA	0.070±0.002	0.224	0.332	0.520	0.685	0.825	0.069±0.005	0.221	0.331	0.521	0.679	0.822
F-G LoRA	0.021±0.002	0.073	0.125	0.248	0.381	0.571	0.019±0.002	0.079	0.139	0.260	0.400	0.582
T-CLIP (Global)	0.071±0.003	0.233	0.349	0.541	0.695	0.837	0.078±0.002	0.242	0.359	0.554	0.710	0.843
T-CLIP (F-G)	0.016±0.002	0.064	0.101	0.194	0.311	0.473	0.012±0.002	0.048	0.084	0.157	0.255	0.394
T-CLIP (Dual)	**0.078±0.004**	**0.252**	**0.374**	**0.571**	**0.729**	**0.862**	**0.084±0.002**	**0.264**	**0.386**	**0.580**	**0.740**	**0.870**
Δ vs. Global LoRA	+10.5%	+12.4%	+12.7%	+9.7%	+6.3%	+4.5%	+21.9%	+19.4%	+16.8%	+11.3%	+9.1%	+5.8%

Table 2: Text-image retrieval recall@K on the FLIR (FLIR Systems, 2025) test set. **Bold**: best result. Underline: second best. F-G = fine-grained. Δ: relative R@K gain of T-CLIP (Dual) over the strongest baseline (Global LoRA). R@1 reported as mean ± std across three independent training runs.

	IMAGE-TO-TEXT RETRIEVAL						TEXT-TO-IMAGE RETRIEVAL					
METHOD	**R@1**	**R@5**	**R@10**	**R@25**	**R@50**	**R@100**	**R@1**	**R@5**	**R@10**	**R@25**	**R@50**	**R@100**
CLIP (Global)	0.014	0.070	0.115	0.242	0.360	0.513	0.014	0.062	0.106	0.229	0.319	0.467
CLIP (F-G)	0.006	0.022	0.039	0.110	0.220	0.332	0.005	0.019	0.032	0.093	0.158	0.253
CLIP-Adapter (Global)	0.045±0.006	0.153	0.247	0.411	0.545	0.700	0.031±0.003	0.130	0.217	0.381	0.529	0.706
CLIP-Adapter (F-G)	0.027±0.003	0.081	0.134	0.250	0.402	0.582	0.019±0.001	0.075	0.133	0.244	0.367	0.555
DeCLIP (Global)	0.049±0.002	0.154	0.228	0.376	0.507	0.642	0.045±0.002	0.158	0.232	0.367	0.510	0.641
DeCLIP (F-G)	0.012±0.001	0.056	0.104	0.197	0.312	0.474	0.015±0.001	0.073	0.114	0.218	0.324	0.474
Global LoRA	0.105±0.003	0.329	0.487	0.678	0.816	0.914	0.106±0.003	0.321	0.459	0.673	0.812	0.907
F-G LoRA	0.027±0.004	0.089	0.150	0.290	0.453	0.642	0.035±0.003	0.127	0.206	0.369	0.528	0.702
T-CLIP (Global)	0.119±0.004	0.353	0.504	0.707	0.833	0.923	0.105±0.007	0.348	0.492	0.712	0.839	0.925
T-CLIP (F-G)	0.025±0.005	0.107	0.182	0.316	0.473	0.649	0.020±0.004	0.069	0.130	0.239	0.361	0.523
T-CLIP (Dual)	**0.123±0.010**	**0.357**	**0.517**	**0.727**	**0.851**	**0.931**	**0.117±0.003**	**0.364**	**0.511**	**0.728**	**0.851**	**0.932**
Δ vs. Global LoRA	+17.1%	+8.5%	+6.2%	+7.2%	+4.3%	+1.9%	+10.4%	+13.4%	+11.3%	+8.2%	+4.8%	+2.8%

dual-caption evaluation achieves the best performance across all metrics, outperforming all single-branch and initialized variants. Notably, initializing Fine-Grained LoRA from Global LoRA weights leads to performance collapse, suggesting that fine-grained thermal representations cannot be reliably obtained by adapting from global ones and that the two optimization trajectories are incompatible. That Global LoRA also fails when initialized from Fine-Grained LoRA validates the geometric incompatibility of the two representation spaces, reflecting the significant divergence reported in Figure 4. Furthermore, Dual LoRA evaluated on global captions (R@1: 0.071) performs comparably to single-branch Global LoRA (R@1: 0.070) while supporting fine-grained retrieval. This capability is entirely absent in any single-branch model, thereby confirming that the dual-branch design retains specialization in each caption domain without requiring separate inference-time models.

Vision vs. Text Encoder Adaptation. Table 5.3 ablates whether LoRA parameters are applied to the vision encoder only, text encoder only, or both. Adapting a single encoder falls significantly short of adapting both jointly, confirming that effective thermal alignment requires simultaneous adaptation of both modalities.

Sensitivity of Retrieval Performance to α. Figure 7 shows retrieval performance across $\alpha \in \{0.0, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ on KAIST. Performance peaks at $\alpha = 0.8$ and degrades at both extremes, with the higher performance at $\alpha = 1.0$ relative to $\alpha = 0.0$ underscoring that global thermal context is the dominant retrieval signal. Nevertheless, the sharp drop at both extremes validates that both caption types

Table 3: Text-image retrieval recall@K on the FMB (Liu et al., 2023) test set. R@1 reported as mean ± std across three independent training runs; all other R@K are means. **Bold**: best result. Underline: second best. F-G = fine-grained. Δ: relative R@K gain of the best T-CLIP variant per metric over Global LoRA.

METHOD	IMAGE-TO-TEXT RETRIEVAL R@1	R@5	R@10	R@25	R@50	R@100	TEXT-TO-IMAGE RETRIEVAL R@1	R@5	R@10	R@25	R@50	R@100
CLIP (Global)	0.043	0.118	0.175	0.318	0.443	0.632	0.018	0.071	0.143	0.243	0.418	0.618
CLIP (F-G)	0.011	0.082	0.143	0.225	0.279	0.532	0.004	0.032	0.068	0.157	0.289	0.479
CLIP-Adapter (Global)	0.075±0.007	0.266	0.409	0.605	0.768	0.905	0.063±0.002	0.259	0.402	0.613	0.775	0.927
CLIP-Adapter (F-G)	0.029±0.004	0.157	0.277	0.455	0.677	0.880	0.039±0.004	0.148	0.243	0.438	0.611	0.863
DeCLIP (Global)	0.095±0.013	0.268	0.400	0.573	0.704	0.834	0.077±0.005	0.277	0.370	0.586	0.700	0.843
DeCLIP (F-G)	0.041±0.009	0.145	0.238	0.402	0.561	0.743	0.038±0.013	0.136	0.239	0.400	0.566	0.754
Global LoRA	0.104±0.015	0.342	0.496	0.731	0.866	0.956	0.100±0.012	0.350	0.510	0.737	0.882	0.960
F-G LoRA	0.037±0.002	0.133	0.248	0.479	0.658	0.842	0.048±0.002	0.185	0.282	0.489	0.666	0.848
T-CLIP (Global)	**0.105±0.012**	**0.362**	**0.539**	0.762	0.871	0.958	0.101±0.006	0.385	0.556	0.787	0.901	0.970
T-CLIP (F-G)	0.042±0.002	0.151	0.245	0.413	0.580	0.788	0.039±0.006	0.144	0.219	0.346	0.526	0.708
T-CLIP (Dual)	0.098±0.017	0.342	0.537	**0.771**	**0.895**	**0.973**	**0.104±0.015**	**0.395**	**0.587**	**0.818**	**0.914**	**0.979**
Δ vs. Global LoRA	+1.0%	+5.8%	+8.7%	+5.5%	+3.3%	+1.8%	+4.0%	+12.9%	+15.1%	+11.0%	+3.6%	+2.0%

Table 4: Ablation of caption types on KAIST (Hwang et al., 2015). Global LoRA and fine-grained (F-G) LoRA are trained and evaluated on their respective caption types. Mixed (50:50) trains a single LoRA on an equal mixture of global and fine-grained captions as a representative mixed training baseline; (Global) and (F-G) denote the evaluation caption type.

CAPTION TYPE	IMAGE-TO-TEXT RETRIEVAL R@1	R@10	R@25	TEXT-TO-IMAGE RETRIEVAL R@1	R@10	R@25
Global LoRA	**0.070**	**0.332**	**0.520**	**0.069**	**0.331**	**0.521**
F-G LoRA	0.021	0.125	0.248	0.019	0.139	0.260
Mixed (50:50) (Global)	0.054	0.318	0.496	0.054	0.287	0.476
Mixed (50:50) (F-G)	0.027	0.160	0.297	0.021	0.152	0.283

are essential; fine-grained thermal captions contribute meaningfully beyond global context alone, and the model benefits from their complementary fusion. We empirically verify that $\alpha = 0.8$ is optimal across all three datasets (KAIST (Hwang et al., 2015), FLIR (FLIR Systems, 2025), and FMB (Liu et al., 2023)), and set it as the default for all reported results.

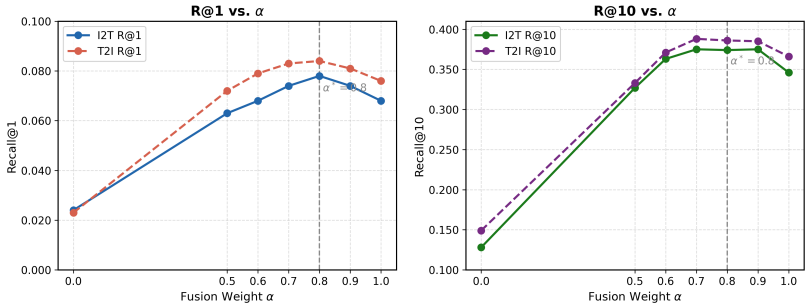

Figure 7: Fusion weight α ablation (KAIST (Hwang et al., 2015)). Peak performance at $\alpha = 0.8$ confirms the complementary contribution of global and fine-grained branches, with steeper degradation toward $\alpha = 0.0$ reflecting the dominance of global thermal context.

Table 5: Ablation of LoRA module designs on KAIST (Hwang et al., 2015). For dual LoRA, the caption type used for evaluation is indicated. A|B denotes that model B is initialized from the weights of model A.

	IMAGE-TO-TEXT RETRIEVAL			TEXT-TO-IMAGE RETRIEVAL		
MODEL	**R@1**	**R@10**	**R@25**	**R@1**	**R@10**	**R@25**
Single LoRA						
Global LoRA	0.070	0.332	0.520	0.069	0.331	0.521
F-G LoRA	0.021	0.125	0.248	0.019	0.139	0.260
Initialized						
F-G LoRA|Global LoRA	0.002	0.013	0.028	0.003	0.021	0.041
Global LoRA|F-G LoRA	0.008	0.044	0.087	0.014	0.076	0.140
Dual LoRA						
Global Captions	0.071	0.349	0.541	0.078	0.359	0.554
F-G Captions	0.016	0.101	0.194	0.012	0.084	0.157
Dual Captions	**0.078**	**0.374**	**0.571**	**0.084**	**0.386**	**0.580**

Table 6: Ablation of LoRA fine-tuning scope across encoder configurations on KAIST (Hwang et al., 2015). All variants use dual captions during evaluation, with $\alpha = 0.8$.

	IMAGE-TO-TEXT RETRIEVAL			TEXT-TO-IMAGE RETRIEVAL		
FINE-TUNING SCOPE	**R@1**	**R@10**	**R@25**	**R@1**	**R@10**	**R@25**
Vision Encoder Only	0.028	0.183	0.328	0.028	0.185	0.318
Text Encoder Only	0.016	0.131	0.237	0.017	0.131	0.257
Both Encoders (T-CLIP Dual)	**0.078**	**0.374**	**0.571**	**0.084**	**0.386**	**0.580**

5.4 Parameter and Training Efficiency

As shown in Table 7, T-CLIP (Dual) achieves the best retrieval performance with two independent 737M LoRA branches which are identical in size to a single Global LoRA. Since the two branches are fully independent, their training is parallelizable, reducing the wall-clock training time to ∼1 hr when run simultaneously on two GPUs, which is equivalent to a single Global LoRA training run. The performance gain is therefore attributable entirely to the decoupled dual-caption design of T-CLIP (Dual) rather than increased model capacity, which is a favorable trade-off between architectural complexity and retrieval performance.

5.5 Cross-Dataset Generalization

Table 8 evaluates T-CLIP (Dual) on the FMB test set when trained on other thermal datasets. Despite the sensor and scene distribution shift, cross-dataset models retain meaningful retrieval capability, with KAIST- and FLIR-trained models recovering 75–85% of FMB-trained performance at R@1. This demonstrates that T-CLIP learns thermal representations that transfer across sensors and environmental conditions, which is a noteworthy result given the high sensor-dependence of thermal imagery. Training on the combined set further improves performance across all metrics, indicating that diversity of thermal scenes and sensor types is beneficial alongside dataset-specific fine-tuning.

5.6 IR-Cap Caption Quality Evaluation

We evaluate IR-Cap caption quality through a human study with five domain-expert annotators assessing 150 stratified image-caption pairs per caption type (300 pairs total), sampled uniformly across dataset origin (KAIST (Hwang et al., 2015), FLIR (FLIR Systems, 2025), FMB (Liu et al., 2023)), time of day, and scene type. Annotators rated each caption on four dimensions for Global captions and five for Fine-Grained captions: Thermal Accuracy (D1), Semantic Correctness (D2), RGB Hallucination (D3, binary), Specificity

Table 7: Efficiency comparison of all methods. Trainable parameters and wall-clock training time are reported for a single training run on KAIST on a single NVIDIA A6000 GPU with batch size 128. ×2 denotes two independent training runs for T-CLIP (Dual), which can be parallelized to ∼1 hr on two GPUs. R@1 results on KAIST (Hwang et al., 2015) test set.

METHOD	TRAINABLE PARAM.	WALL-CLOCK TIME	I2T R@1	T2I R@1
Zero-shot CLIP	0	0	0.003	0.002
CLIP-Adapter	131M	∼40 min	0.018	0.020
LoRA (Text Only)	294M	∼1 hr	0.016	0.017
LoRA (Vision Only)	442M	∼1 hr	0.028	0.028
Global LoRA	737M	∼1 hr	0.070	0.069
T-CLIP (Dual)	**737M × 2**	**∼1 hr × 2**	**0.078**	**0.084**

Table 8: Cross-dataset generalization on the FMB (Liu et al., 2023) test set. Each row denotes the dataset used for training. Combined set denotes training on all three datasets, KAIST (Hwang et al., 2015), FLIR (FLIR Systems, 2025), and FMB (Liu et al., 2023), evaluated on FMB.

	IMAGE-TO-TEXT RETRIEVAL			TEXT-TO-IMAGE RETRIEVAL		
TRAINED ON	R@1	R@10	R@25	R@1	R@10	R@25
FMB	0.098	0.537	0.771	0.104	0.587	0.818
KAIST	0.079	0.421	0.671	0.082	0.496	0.668
FLIR	0.075	0.375	0.561	0.111	0.439	0.604
Combined Set	**0.132**	**0.654**	**0.850**	**0.125**	**0.650**	**0.864**

(D4), and Emissivity Reasoning (D5, Fine-Grained only). A caption is approved if it meets minimum thresholds on all applicable dimensions simultaneously. Full protocol details, annotator assignment design, rubric anchors, and failure case analysis are provided in section A.3 of the appendix.

Table 9 presents the results. Global Thermal Captions achieve an overall approval rate of 97.3% [94.1, 99.0], while Fine-Grained Thermal Captions achieve 84.0% [77.8, 89.0] (95% CI). All dimensions achieve Krippendorff's $\alpha > 0.60$ (substantial agreement), with RGB Hallucination reaching $\alpha > 0.80$, consistent with its binary and objective nature. Global captions score significantly higher on Thermal Accuracy (D1: 3.58 vs. 3.21, paired t-test $p<0.001$), indicating that scene-level thermal properties are more reliably inferred from RGB anchors than object-level heat signatures. The primary limiting factor for Fine-Grained captions is Emissivity Reasoning (D5: 89.3% pass rate), reflecting object-level thermal physics are not completely estimated from visible-spectrum appearance, a limitation of the RGB-anchoring strategy. Failure case analysis is provided in appendix section A.3.4.

5.7 Applying T-CLIP to Thermal Image Generation

CLIP is widely used in many downstream tasks, including segmentation, detection, and text-to-image generation models like Stable Diffusion. However, standard CLIP struggles with thermal domain understanding because it was trained exclusively on RGB images. We explore whether T-CLIP can serve as a plug-and-play replacement for the text encoder in SDXL (Podell et al., 2023a). Specifically, we replace the original CLIP ViT-L/14 text encoder in SDXL with our T-CLIP text encoder. Because SDXL's UNet was trained only on RGB images and cannot directly generate thermal outputs, we fine-tune the UNet on KAIST thermal image–caption pairs to adapt its decoding capabilities to the thermal domain.

[2]Pass thresholds: $D1, D2 \geq 3$; $D3 = 0$; $D4, D5 \geq 2$. Krippendorff's $\alpha>0.60$ on all dimensions; $\alpha>0.80$ for D3.

Table 9: IR-Cap human evaluation results (n=150 per caption type, stratified across KAIST (Hwang et al., 2015), FLIR (FLIR Systems, 2025), and FMB (Liu et al., 2023)). Scores are mean ± std on the 1–4 Likert scale (higher is better); D3 is hallucination-free percentage. Appr.% denotes the percentage of captions meeting the per-dimension pass threshold. Overall approval requires passing all applicable dimensions simultaneously; 95% CIs via Wilson score interval. †Fine-grained only. [2]

	GLOBAL		FINE-GRAINED	
DIMENSION	**SCORE**	**APPR.%**	**SCORE**	**APPR.%**
D1: Thermal Accuracy	3.58 ± 0.52	98.7%	3.21 ± 0.68	91.3%
D2: Semantic Correctness	3.74 ± 0.48	99.3%	3.48 ± 0.59	96.0%
D3: RGB Hallucination-free	98.0%	—	93.3%	—
D4: Specificity	3.42 ± 0.61	99.3%	3.35 ± 0.64	97.3%
D5: Emissivity†	—	—	2.98 ± 0.74	89.3%
Overall Approval	**97.3**% [94.1, 99.0]		**84.0**% [77.8, 89.0]	

Table 10: User study results for thermal image generation (30 pairs, 5 annotators, 150 ratings per dimension). SDXL: zero-shot Stable Diffusion XL (Podell et al., 2023b) with standard CLIP text encoder. T-CLIP: our thermally adapted text encoder replacing the standard CLIP encoder in SDXL. Δ = T-CLIP score − SDXL score (mean difference across 150 ratings). †Sign test used because SDXL variance = 0 (unanimous score of 1 on D1 and D2 precludes Wilcoxon). ‡n.s. = not significant ($p > 0.05$ after Benjamini-Hochberg correction across three comparisons, FDR = 0.05).[4]

DIMENSION	**SDXL**	**T-CLIP**	**Δ**	**SIG.**
D1: Thermal plausibility	1.00 ± 0.00	3.98 ± 0.14	+2.98	$p \ll 0.001^{\dagger}$
D2: Physics correctness	1.00 ± 0.00	3.0 ± 0.5	+2.0	$p \ll 0.001^{\dagger}$
D3: Prompt faithfulness	2.9 ± 0.4	3.0 ± 0.5	+0.1	n.s.‡
Preference	100% T-CLIP preferred ($n = 150$, $p \ll 0.001$)			

Qualitatively, as shown in Figure 8, T-CLIP generates physically plausible thermal phenomena — vehicle heat emissions, pedestrian body heat, and atmospheric scattering, under challenging conditions including night scenes and foggy environments. Additional generated samples across different illumination and weather conditions are presented in section A.5 of the appendix.

To evaluate thermal generation quality, we conducted a user study with five annotators with domain expertise on 30 prompt pairs spanning five scene categories (full protocol and other calculations are detailed in appendix section A.5). Annotators rated each generated image on Thermal Plausibility (D1), Physics Correctness (D2 – a binary checklist based on thermal imaging principles, e.g.: *warm* → *brighter; cool* → *darker*), and Prompt Faithfulness (D3), each on a 1–4 scale.

As shown in Table 10, T-CLIP + SDXL achieves excellent ratings in terms of Thermal Plausibility (3.98±0.14) and substantially improved Physics Correctness (3.0 ± 0.5), with zero-shot SDXL receiving a score of 1 on both D1 and D2, suggesting that all generated images are pseudo-coloured RGB with no visible thermal characteristics. Prompt Faithfulness is equivalent across both models ($\Delta = +0.1$, $p > 0.05$, n.s.), suggesting that T-CLIP adds thermal quality without altering scene content generation. T-CLIP+SDXL was preferred across all 150 pairwise comparisons (100%, $p \ll 0.001$). Our user study demonstrates that T-CLIP's thermal representations transfer meaningful thermal reasoning capabilities to existing generation frameworks without requiring architectural changes.

[4]Pass thresholds and statistical notes: †Sign test used because SDXL variance = 0. ‡n.s. = not significant ($p > 0.05$, Benjamini-Hochberg corrected, FDR = 0.05).

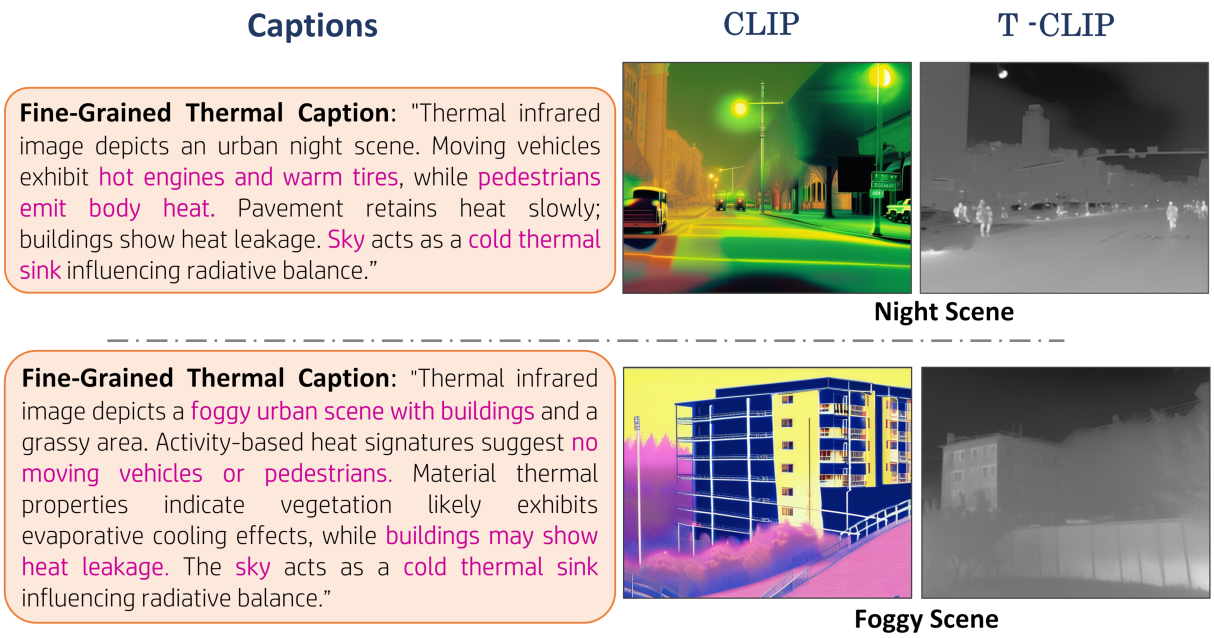

Figure 8: T-CLIP as a plug-and-play replacement for the text encoder in SDXL for thermal image generation. SDXL equipped with the standard CLIP encoder produces pseudo-coloured RGB-like outputs lacking meaningful thermal characteristics. In contrast, replacing the CLIP encoder with T-CLIP enables SDXL to generate physically plausible thermal images that capture pedestrian body heat, vehicle heat emissions in nighttime scenes, and building heat leakage under foggy conditions.

6 Conclusion, Limitations, and Future Work

We introduced IR-Cap, the first physics-aware thermal captioning pipeline and dataset providing complementary global and fine-grained thermal descriptions across three public benchmarks, and T-CLIP, a decoupled dual-LoRA framework for thermal CLIP adaptation. The central insight of this work is that global scene-level thermal context and object-level heat signatures are geometrically incompatible and cannot be meaningfully optimized within a shared embedding space. This finding, validated through feature geometry analysis and controlled ablations, directly motivates our decoupled design: two independent specialized branches trained separately on IR-Cap's global and fine-grained captions, then fused at inference. T-CLIP achieves consistent improvements across three thermal benchmarks on cross-modal retrieval and offers preliminary evidence of its transferability to text-conditioned thermal image generation. We release IR-Cap and our captioning pipeline to facilitate future progress in this area.

Limitations and Future Work. IR-Cap relies on paired RGB images as semantic anchors for generating thermal captions, which limits the modelling of fine-grained thermal phenomena not directly observable in the visible spectrum. This motivates future work on richer thermal metadata annotation strategies to enable more precise characterization of object-level heat signatures. While T-CLIP can be used for text-conditioned thermal image generation as a plug-and-play text encoder replacement, a systematic study of thermal image generation constitutes a substantial independent research direction beyond the scope of this work.

References

Muhammad Ali Farooq, Waseem Shariff, David O'callaghan, Arcangelo Merla, and Peter Corcoran. On the role of thermal imaging in automotive applications: A critical review. *IEEE Access*, 2023.

FLIR Systems. FLIR ADAS Dataset. https://oem.flir.com/en-in/solutions/automotive/adas-dataset-form/, 2025. Accessed: 2026-02-20.

Kevin Frans, Lisa B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *NeurIPS*, 2022.

Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: a survey. *Machine vision and applications*, 25:245–262, 2014.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*. OpenReview.net, 2022.

Yunze He, Baoyuan Deng, Hongjin Wang, Liang Cheng, Ke Zhou, Siyuan Cai, and Francesco Ciampa. Infrared machine vision and infrared thermography with deep learning: A review. *Infrared physics & technology*, 116:103754, 2021.

I Hirsh, L Shkedy, D Chen, N Fishler, Y Hagbi, A Koifman, Y Openhaim, I Vaserman, M Singer, and I Shtrichman. Hybrid dual-color mwir detector for airborne missile warning systems. In *Infrared Technology and Applications XXXVIII*, volume 8353, pp. 189–200. SPIE, 2012.

Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1037–1045, 2015.

Y. Iwashita, K. Nakashima, S. Rafol, A. Stoica, and R. Kurazume. Mu-net: Deep learning-based thermal ir image estimation from rgb image. In *Unpublished*, 2023.

VV Kniaz, VS Gorbatsevich, and VA Mizginov. Thermalnet: a deep convolutional network for synthetic thermal image generation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:41–45, 2017.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123 (1):32–73, 2017.

Zhengfeng Lai, Zhuoheng Li, Luca Cerny Oliveira, Joohi Chauhan, Brittany N Dugger, and Chen-Nee Chuah. Clipath: Fine-tune clip with visual feature fusion for pathology image analysis towards minimizing data collection efforts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2374–2380, 2023.

Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*. OpenReview.net, 2022a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. *CoRR*, abs/2112.03857, 2021.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*. OpenReview.net, 2022b.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *ECCV*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8115–8124, 2023.

Christian Mayr, Christian Kubler, Norbert Haala, and Michael Teutsch. Narrowing the synthetic-to-real gap for thermal infrared semantic image segmentation using diffusion-based conditional image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3131–3141, 2024.

Tran Xuan Bach Nguyen, Kent Rosser, and Javaan Chahl. A review of modern thermal imaging sensor technology and applications for autonomous aerial navigation. *Journal of Imaging*, 7(10):217, 2021.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023a.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023b.

Tayeba Qazi, Brejesh Lall, and Prerana Mukherjee. Thermaldiff: A diffusion architecture for thermal image synthesis. *Journal of Visual Communication and Image Representation*, 111:104524, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022.

Lingyan Ran, Lidong Wang, Guangcong Wang, Peng Wang, and Yanning Zhang. Diffv2ir: visible-to-infrared diffusion model via vision-language understanding. *arXiv preprint arXiv:2503.19012*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *NeruIPS*, 2022.

Michael Vollmer and Klaus-Peter Möllmann. *Infrared thermal imaging: fundamentals, research and applications*. John Wiley & Sons, 2018.

AN Wilson, Khushi Gupta, Balu Harshavardan Koduru, Abhinav Kumar, Ajit Jha, and Linga Reddy Cenkeramaddi. Recent advances in thermal imaging and its applications using machine learning: A review. *IEEE Sensors Journal*, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pp. 18113–18123. IEEE, 2022.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6757–6767, 2023.

Table 11: Pairwise Welch's t-test results for Figure 1 (main manuscript) image-text cosine similarity distributions on the KAIST (Hwang et al., 2015) test set. Benjamini-Hochberg (BH) correction applied across all three comparisons (FDR = 0.05). n=2,252 image-caption pairs per method. p-values are unchanged in practical terms after correction, confirming the negligible effect of adjustment given effect sizes of this magnitude.[6]

COMPARISON	MEAN A	MEAN B	Δ	t-STAT	p_{RAW}	p_{BH}
Zero-shot CLIP vs Global LoRA	0.3449	0.3322	−0.0127	11.559	2.83×10^{-30}	2.83×10^{-30}
Global LoRA vs T-CLIP	0.3322	0.3716	+0.0394	−30.350	1.86×10^{-183}	5.58×10^{-183}
Zero-shot CLIP vs T-CLIP	0.3449	0.3716	+0.0267	−29.132	7.61×10^{-167}	1.14×10^{-166}

Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International conference on image processing (ICIP)*, pp. 276–280. IEEE, 2020.

A Appendix

A.1 Quantifying the Thermal Perception Gap

Table 11 reports the full pairwise Welch's t-test results for the image-text cosine similarity distributions shown in sections 1 and 4.1, including BH-corrected p-values. All three pairwise differences are statistically significant with $p \ll 0.001$ before and after correction, confirming the robustness of the reported cosine similarity comparisons.

A.2 IR-Cap Dataset — Instruction Prompts

As discussed in section 3 we employed a dual prompting strategy with Qwen2.5-VL-72B-Instruct, using the visible-spectrum images as semantic context for generating captions for corresponding thermal images. Figure 9 shows the two instruction prompts used for Global and Fine-Grained caption generation respectively.

A.3 IR-Cap Human Evaluation

This section provides complete protocol details for the human evaluation of the IR-Cap dataset, supporting section 5.6. It includes the full rubric, stratified sampling design, annotator assignment matrix, failure case analysis, and inter-annotator agreement.

A.3.1 Full Rubric Anchor Descriptions

Table 12 provide complete anchor descriptions and examples for all rating dimensions. Annotators were provided this table as a printed reference sheet. D3 is binary (0/1); D5 applies to Fine-Grained captions only. Pass thresholds: $\text{D1}, \text{D2} \geq 3$; $\text{D3} = 0$; $\text{D4}, \text{D5} \geq 2$.

Approval rate computation. The per-dimension approval rate (Appr.%) for dimension D_i is:

$$\text{Appr.}_{D_i} = \frac{|\{c : s_{D_i}(c) \geq \tau_i\}|}{n}, \tag{7}$$

where $s_{D_i}(c)$ is the mean annotator score for caption c, τ_i is the pass threshold, and n=150. Overall Approval requires passing *all applicable* dimensions simultaneously:

$$\text{Overall Appr.} = \frac{|\{c : \forall\, i \in \mathcal{D}(c),\ s_{D_i}(c) \geq \tau_i\}|}{n}, \tag{8}$$

[6]Method means (cosine similarity): Zero-shot CLIP = 0.3449±0.0203; Global LoRA = 0.3322±0.0482; T-CLIP = 0.3716±0.0384. Δ= Mean B − Mean A. Two-sided Welch's t-test (unequal variance).

Instruction Prompt 1:

```
""" Provide a caption for the image using a maximum of 77 tokens.
Describe the scene in a broad and objective manner, explicitly
mentioning all visible objects, activities, lighting, weather
conditions and time of day. Avoid using emotional or subjective
terms.
"""
```

Instruction Prompt 2:

```
"""Analyze RGB images to predict thermal infrared characteristics. For
   each image, generate exactly one caption that:

1. Begins with "Thermal infrared image"
2. Briefly describes the scene content
3. Analyzes thermal properties based on:
   - Time of day, weather conditions, and environmental factors
   - ACTIVITY-BASED HEAT SIGNATURES:
     - Moving vehicles: Hot engines, warm tires, exhaust patterns
     - Parked vehicles: Cooling engines, ambient temperature alignment
     - Pedestrians: Body heat emissions, movement-generated warmth
     - Recent activity: Warm brakes, recently parked vehicles
   - MATERIAL THERMAL PROPERTIES:
     - Metals: Rapid temperature changes (engines, exhaust)
     - Pavement: Heat retention, slow temperature changes
     - Vegetation: Evaporative cooling, typically cooler
     - Buildings: Heat leakage through windows/walls
     - Sky: Cold thermal sink affecting radiative balance
4. Uses objective, scientific language without colors or emotional terms
5. Strictly stick to caption length of 77-token CLIP models.
"""
```

Figure 9: Instruction prompts for IR-Cap caption generation pipeline. Instruction Prompt 1 generates Global Thermal Captions describing scene-level environmental context. Instruction Prompt 2 generates Fine-Grained Thermal Captions encoding object-level heat signatures and thermal phenomena.

where $\mathcal{D}(c)$ denotes the applicable dimensions for caption c (D1–D4 for Global; D1–D5 for Fine-Grained). Overall Approval is at most the minimum per-dimension approval rate and is not the average of per-dimension rates.

A.3.2 Stratified Sampling Design

Samples were drawn uniformly across dataset origin, time of day, and scene type; Table 13 details the breakdown.

A.3.3 Annotator Assignment Matrix

Each annotator evaluates 63 items per caption type (126 total), comprising 20 shared overlap items and approximately 43 unique items per batch, yielding an average of 3.3 ratings per item. For items where annotator scores differed by more than one point on any dimension, a sixth expert adjudicator provided a tie-breaking score (7 items, 2.3% of overlap items).

A.3.4 Failure Case Analysis

Table 15 summarises the three systematic failure categories identified in the 24 Fine-Grained captions failing overall approval, all arising from thermal phenomena not directly observable in the RGB semantic anchor.

A.3.5 Pairwise Annotator Agreement

All five annotators achieved substantial pairwise agreement (Krippendorff's α>0.60) across all dimensions on the 20 overlap items, consistent with the overall α values reported in the main paper. D3 (RGB Hallucination) consistently achieved the highest pairwise agreement (α>0.80), reflecting both its binary scale and the straightforward nature of the judgment: annotators simply check whether the caption contains explicit colour or texture terms (e.g., "white", "yellow", "dark") that are meaningless in the thermal infrared domain.

Data Availability. All caption pairs and images used for the human evaluation of the IR-Cap dataset are available at `https://drive.google.com/drive/folders/10Xt5lVwSrkiSvJdmFRFMBZtg0SRUvhhS?usp=drive_link`.

A.4 Retrieval Results with Multi-Seed Statistics

Tables 16, 17 and 18 report the full Recall@K retrieval results across all three benchmarks. R@K is reported as mean $\pm$ std across three independent training runs (seeds 0, 42, 123). Zero-shot CLIP results are deterministic and reported as point estimates only.

A.5 Thermal Image Generation: User Study, Quantitative Metrics, and Additional Samples

This appendix provides complete protocol details supporting the user study of generated thermal images presented in Figure 8, 11 and 12. section A.5.1 describes the user study protocol, Table 19 gives the full D2 Physics Correctness checklist, section A.5.3 gives statistical analysis notes, Figure 10 reports FID and CLIP Score of the generated thermal images for reference, and Figures 11 and 12 present additional generated samples.

A.5.1 Study Protocol

Method identity was concealed throughout: annotators saw only "Image A" and "Image B", with A/B labels randomised independently per annotator per pair using a fixed per-annotator seed to prevent position bias. The full text prompts were displayed above both images for each pair. Three practice pairs (excluded from analysis) preceded the main study. A private mapping file recorded the A/B assignment per annotator per pair; post-hoc, each A/B response was converted to a method-level score as $s^{\text{tclip}}_{a,p,d} = r^{A}_{a,p,d}$ if Image A was T-CLIP, else $r^{B}_{a,p,d}$, yielding n=150 paired scores per dimension (5 annotators $\times$ 30 pairs).

A.5.2 D2 Physics Correctness Checklist

Scoring formula. For each image i and annotator a, let $c_{a,i,j} \in \{0, 1, \text{N/A}\}$ denote the response to checklist item j. The per-image D2 score is:

$$s^{D2}_{a,i} = \left\lfloor \frac{\sum_{j:\, c_{a,i,j} \neq \text{N/A}} c_{a,i,j}}{|\{j : c_{a,i,j} \neq \text{N/A}\}|} \times 3 \right\rfloor + 1, \tag{9}$$

clamped to $[1, 4]$, where $\lfloor\cdot\rceil$ denotes rounding to the nearest integer. N/A items are excluded from both numerator and denominator. If all items are N/A the pair is excluded from D2 analysis.
Refer to Table 19 for the detailed checklist for physics correctness.

A.5.3 Statistical Analysis

Statistical tests. D1 (Thermal Plausibility) and D2 (Physics Correctness) are assessed via the sign test rather than the Wilcoxon signed-rank test, because zero-shot SDXL received a unanimous score of 1 on both dimensions across all 150 ratings, yielding zero variance; the sign test does not require non-zero differences ($p \approx 4.9 \times 10^{-46}$ for both). D3 (Prompt Faithfulness) is assessed via the Wilcoxon signed-rank test; the difference was not significant ($p > 0.05$), consistent with both models generating correct scene content from the prompt. Benjamini-Hochberg correction (FDR $= 0.05$) is applied across all three pairwise comparisons to control the false discovery rate. Annotator preference (T-CLIP + SDXL vs. zero-shot SDXL) is assessed via a one-sided binomial test (H_0: preference probability $= 0.5$); all 150 decisive comparisons favoured T-CLIP + SDXL, yielding $p \approx 4.9 \times 10^{-46}$, as reported in Table 10. of ,aim manuscript.

Score aggregation. For D1 and D3, each annotator assigns a Likert score $s_{a,p,d} \in \{1, 2, 3, 4\}$ per pair. For D2, scores are derived from a per-image checklist via equation 9. In all cases, the reported mean and standard deviation are computed over all n=150 paired scores (5 annotators $\times$ 30 pairs) as follows:

$$\mu_d = \frac{1}{150} \sum_{a,p} s_{a,p,d}, \qquad \sigma_d = \mathrm{std}\big(\{s_{a,p,d}\}_{a,p}\big), \tag{10}$$

where $s_{a,p,d}$ is the score assigned by annotator a to pair p on dimension d. The method-level mean difference on dimension d is $\Delta = \mu_d^{\text{T-CLIP}} - \mu_d^{\text{SDXL}}$.

A.5.4 FID and CLIP Score

Figure 10 reports FID and CLIP Score for completeness, providing additional quantitative evaluation of the generated thermal images beyond the user study. Both metrics rely on RGB-trained features and are fundamentally misaligned with the thermal domain; the user study (Section 5.6 of the main manuscript) provides the primary evaluation. The consistent FID improvements (+14% to +55%) and CLIP Score improvements (+6% to +19%) across all three datasets are nonetheless directionally supportive.

Data Availability. All 30 prompts and corresponding generated image pairs used for the user study of thermal image generation are available at `https://drive.google.com/drive/folders/1mxPTpRJgI54XvG5Je8LTEci9PlslPYSK?usp=drive_link`.

A.5.5 Additional Generated Samples

We provide additional qualitative examples of thermal images generated using T-CLIP with SDXL using a variety of text prompts in Figures 11 and 12.

A.6 Implementation, Training, and Hyperparameter Details

All methods use CLIP (Radford et al., 2021) ViT-Base-Patch16 as the backbone and are trained on thermal image–caption pairs generated by the IR-Cap pipeline. Table 20 lists hyperparameters shared across all methods; Table 21 lists method-specific configurations. The T-CLIP training and inference pseudocode is shown in Figure 13.

A.6.1 Shared Training Configuration

Table 20 lists the implementation details and shared training configurations used across all experiments. All experiments were performed on a single NVIDIA RTX A6000 GPU with 48 GB of memory.

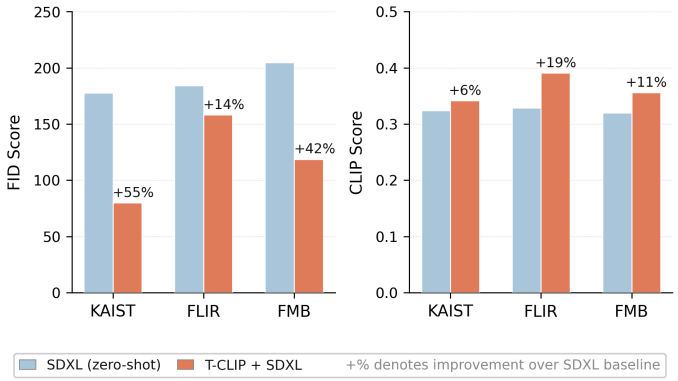

Figure 10: FID (lower is better) and CLIP Score (higher is better) for zero-shot SDXL vs T-CLIP + SDXL across KAIST (Hwang et al., 2015), FLIR (FLIR Systems, 2025), and FMB (Liu et al., 2023).

A.6.2 Method-Specific Configuration

Table 21 summarises the configuration parameters unique to each method. LoRA (Global) and LoRA (F-G) share identical hyperparameters, differing only in the caption type used for training. CLIP-Adapter uses a residual bottleneck adapter on top of frozen CLIP features, controlled by the blend ratio α_{adp}. DeCLIP fine-tunes all CLIP parameters with an additional image–image and text–text nearest-neighbour consistency loss computed against a momentum encoder (EMA update), using a lower learning rate. All methods are trained for an identical number of optimization steps (10,000) to ensure a fair comparison.

A.6.3 T-CLIP Training and Inference Pseudocode

Figure 13 illustrates the T-CLIP training and inference procedure. The Global LoRA (θ_g) and Fine-Grained LoRA (θ_f) branches are optimised independently on their respective caption types with no shared gradient flow, and fused at inference via ensemble weight α=0.8.

Table 12: Full anchor descriptions for D1 (thermal accuracy), D2 (semantic correctness), D3 (RGB hallucination), D4 (specificity), and D5 (emissivity reasoning). Pass thresholds: D1, D2 $\geq$ 3; D3 = 0; D4, D5 $\geq$ 2. D5 applies to fine-grained captions only. Overall approval requires passing all applicable dimensions simultaneously.

DIM.	SCORE	DESCRIPTION	EXAMPLE
D1: Thermal accuracy (1–4, pass $\geq$ 3)			
D1	1	Physically impossible thermal claims	"The sky radiates heat downward, warming the wet pavement"
	2	Thermally neutral; no physics reasoning	"People and cars are visible on a road at night"
	3	Correct thermal reasoning; incomplete	"Scene shows warm objects against cool background at night"
	4	Full correct thermal reasoning	"Night scene: reduced ambient temperature; wet pavement retains residual heat; overcast sky acts as cold thermal sink"
D2: Semantic correctness (1–4, pass $\geq$ 3)			
D2	1	Wrong scene described entirely	Caption describes indoor scene; image is outdoor road
	2	Correct environment; wrong objects	"Cars on highway" but image shows pedestrians
	3	Correct objects; minor attribute errors	"Two pedestrians" but image shows three
	4	All objects and scene correctly identified	—
D3: RGB hallucination (binary 0/1, pass = 0)			
D3	0	No RGB-style colour or texture language used	—
	1	RGB hallucination detected	"White lane markings"; "Yellow streetlights"; "Dark coloured car"
D4: Specificity (1–4, pass $\geq$ 2)			
D4	1	Completely generic	"An infrared thermal image"
	2	Scene identified; no thermal detail	"People walking on a road at night"
	3	Some thermal context present	"Night scene with warm pedestrians and cooler background"
	4	Rich environmental thermal detail	"Late evening urban scene; ambient temperature low; wet road retains residual heat; buildings show heat leakage"
D5: Emissivity reasoning (1–4, pass $\geq$ 2) — fine-grained only			
D5	1	No object-level thermal reasoning	"Thermal image of people and cars"
	2	Generic warm/cool labels only	"People are warm, cars are cool"
	3	Material-specific reasoning; incomplete	"Moving vehicles show warm engines; pedestrians emit body heat from torso"
	4	Full emissivity and physics reasoning	"Moving vehicles exhibit hot engine blocks and warm tires from friction; pedestrians emit body heat at torso; vegetation appears cool due to evaporative cooling"

Table 13: Stratified sampling design across KAIST (Hwang et al., 2015), FLIR (FLIR Systems, 2025), and FMB (Liu et al., 2023), time of day, and scene type.

CAPTION TYPE	DATASET	DAY	NIGHT	OTHER	TOTAL
Global (150)	KAIST	17	17	16	50
	FLIR	17	17	16	50
	FMB	17	17	16	50
	Total	51	51	48	150
Fine-grained (150)	KAIST	17	17	16	50
	FLIR	17	17	16	50
	FMB	17	17	16	50
	Total	51	51	48	150
Grand total		102	102	96	300

Table 14: Annotator assignment matrix, applied independently to both global and fine-grained sets. ✓ = assigned; — = not assigned.

ITEM SET	ANN. 1	ANN. 2	ANN. 3	ANN. 4	ANN. 5	ITEMS
Overlap (items 1–20)	✓	✓	✓	✓	✓	20
Batch A (items 21–63)	✓	✓	✓	—	—	43
Batch B (items 64–107)	—	✓	✓	✓	—	44
Batch C (items 108–150)	—	—	✓	✓	✓	43
Per annotator	63	63	63	63	63	

Table 15: Representative limitation cases in fine-grained IR-Cap captions arising from the RGB-anchoring strategy. Percentage of failures (24 total) per category shown in parentheses.

CATEGORY (SHARE)	GENERATED CAPTION	IDEAL CAPTION	ROOT CAUSE
State ambiguity (54%)	"Vehicles exhibit warm engines and hot tires from friction"	"Moving vehicles exhibit hot engine blocks; stationary vehicles show cooling engines with reduced heat signatures"	Motion state not inferable from RGB; VLMs cannot map motion to thermal signatures
Intra-object distribution (29%)	"Pedestrians emit body heat"	"Torso and head show strong heat emission; extremities cooler; clothing acts as partial thermal insulation"	Within-object thermal gradients not inferable in RGB
Emissivity mechanism (17%)	"Wet road retains heat from daytime solar radiation"	"Wet surface shows elevated emissivity (≈0.97 vs. dry ≈0.92), appearing thermally brighter independent of actual temperature"	Emissivity physics concept absent from VLM training distribution

Table 16: Text-image retrieval recall@K on the KAIST (Hwang et al., 2015) test set. All values reported as mean ± std across three independent training runs. **Bold**: best result. Underline: second best. F-G = fine-grained. Δ: relative R@K gain of T-CLIP (Dual) over the strongest baseline (Global LoRA).

IMAGE-TO-TEXT RETRIEVAL						
METHOD	**R@1**	**R@5**	**R@10**	**R@25**	**R@50**	**R@100**
CLIP (Global)	0.003	0.016	0.034	0.068	0.110	0.179
CLIP (F-G)	0.001	0.004	0.008	0.029	0.057	0.096
CLIP-Adapter (Global)	0.018±0.001	0.072±0.004	0.123±0.002	0.229±0.003	0.357±0.003	0.521±0.002
CLIP-Adapter (F-G)	0.009±0.002	0.041±0.002	0.081±0.005	0.167±0.006	0.283±0.005	0.436±0.006
DeCLIP (Global)	0.038±0.002	0.129±0.001	0.206±0.009	0.349±0.005	0.497±0.008	0.644±0.004
DeCLIP (F-G)	0.013±0.000	0.048±0.002	0.088±0.002	0.169±0.002	0.273±0.008	0.433±0.013
Global LoRA	0.070±0.002	0.224±0.004	0.332±0.013	0.520±0.007	0.685±0.015	0.825±0.013
F-G LoRA	0.021±0.002	0.073±0.007	0.125±0.005	0.248±0.010	0.381±0.006	0.571±0.003
T-CLIP (Global)	0.071±0.003	0.233±0.002	0.349±0.003	0.541±0.004	0.695±0.006	0.837±0.011
T-CLIP (F-G)	0.016±0.002	0.064±0.003	0.101±0.007	0.194±0.007	0.311±0.010	0.473±0.015
T-CLIP (Dual)	**0.078±0.004**	**0.252±0.003**	**0.374±0.004**	**0.571±0.008**	**0.729±0.013**	**0.862±0.009**
Δ vs. Global LoRA	+10.5%	+12.4%	+12.7%	+9.7%	+6.3%	+4.5%

TEXT-TO-IMAGE RETRIEVAL						
METHOD	**R@1**	**R@5**	**R@10**	**R@25**	**R@50**	**R@100**
CLIP (Global)	0.002	0.013	0.027	0.048	0.080	0.135
CLIP (F-G)	0.001	0.007	0.014	0.030	0.048	0.088
CLIP-Adapter (Global)	0.020±0.002	0.070±0.002	0.119±0.003	0.234±0.006	0.360±0.001	0.538±0.003
CLIP-Adapter (F-G)	0.010±0.000	0.046±0.000	0.082±0.003	0.166±0.004	0.281±0.008	0.445±0.004
DeCLIP (Global)	0.041±0.000	0.133±0.004	0.209±0.002	0.357±0.005	0.496±0.014	0.650±0.005
DeCLIP (F-G)	0.014±0.000	0.056±0.002	0.094±0.001	0.175±0.006	0.285±0.013	0.439±0.010
Global LoRA	0.069±0.005	0.221±0.008	0.331±0.010	0.521±0.013	0.679±0.012	0.822±0.010
F-G LoRA	0.019±0.002	0.079±0.010	0.139±0.008	0.260±0.002	0.400±0.007	0.582±0.012
T-CLIP (Global)	0.078±0.002	0.242±0.004	0.359±0.009	0.554±0.008	0.710±0.010	0.843±0.009
T-CLIP (F-G)	0.012±0.002	0.048±0.004	0.084±0.006	0.157±0.010	0.255±0.012	0.394±0.018
T-CLIP (Dual)	**0.084±0.002**	**0.264±0.005**	**0.386±0.010**	**0.580±0.008**	**0.740±0.007**	**0.870±0.011**
Δ vs. Global LoRA	+21.9%	+19.4%	+16.8%	+11.3%	+9.1%	+5.8%

Table 17: Text-image retrieval recall@K on the FLIR (FLIR Systems, 2025) test set. All values reported as mean ± std across three independent training runs. **Bold**: best result. Underline: second best. F-G = fine-grained. Δ: relative R@K gain of T-CLIP (Dual) over the strongest baseline (Global LoRA).

IMAGE-TO-TEXT RETRIEVAL						
METHOD	**R@1**	**R@5**	**R@10**	**R@25**	**R@50**	**R@100**
CLIP (Global)	0.014	0.070	0.115	0.242	0.360	0.513
CLIP (F-G)	0.006	0.022	0.039	0.110	0.220	0.332
CLIP-Adapter (Global)	0.045±0.006	0.153±0.005	0.247±0.010	0.411±0.016	0.545±0.014	0.700±0.014
CLIP-Adapter (F-G)	0.027±0.003	0.081±0.001	0.134±0.003	0.250±0.005	0.402±0.012	0.582±0.005
DeCLIP (Global)	0.049±0.002	0.154±0.002	0.228±0.005	0.376±0.005	0.507±0.002	0.642±0.004
DeCLIP (F-G)	0.012±0.001	0.056±0.003	0.104±0.008	0.197±0.002	0.312±0.001	0.474±0.010
Global LoRA	0.105±0.003	0.329±0.011	0.487±0.008	0.678±0.003	0.816±0.004	0.914±0.002
F-G LoRA	0.027±0.004	0.089±0.007	0.150±0.008	0.290±0.016	0.453±0.016	0.642±0.014
T-CLIP (Global)	0.119±0.004	0.353±0.007	0.504±0.013	0.707±0.007	0.833±0.007	0.923±0.003
T-CLIP (F-G)	0.025±0.005	0.107±0.006	0.182±0.004	0.316±0.009	0.473±0.032	0.649±0.014
T-CLIP (Dual)	**0.123±0.010**	**0.357±0.006**	**0.517±0.015**	**0.727±0.005**	**0.851±0.005**	**0.931±0.008**
Δ vs. Global LoRA	+17.1%	+8.5%	+6.2%	+7.2%	+4.3%	+1.9%

TEXT-TO-IMAGE RETRIEVAL						
METHOD	**R@1**	**R@5**	**R@10**	**R@25**	**R@50**	**R@100**
CLIP (Global)	0.014	0.062	0.106	0.229	0.319	0.467
CLIP (F-G)	0.005	0.019	0.032	0.093	0.158	0.253
CLIP-Adapter (Global)	0.031±0.003	0.130±0.007	0.217±0.009	0.381±0.024	0.529±0.024	0.706±0.017
CLIP-Adapter (F-G)	0.019±0.001	0.075±0.006	0.133±0.011	0.244±0.005	0.367±0.009	0.555±0.009
DeCLIP (Global)	0.045±0.002	0.158±0.011	0.232±0.001	0.367±0.001	0.510±0.008	0.641±0.005
DeCLIP (F-G)	0.015±0.001	0.073±0.003	0.114±0.008	0.218±0.012	0.324±0.010	0.474±0.001
Global LoRA	0.106±0.003	0.321±0.007	0.459±0.006	0.673±0.006	0.812±0.007	0.907±0.006
F-G LoRA	0.035±0.003	0.127±0.005	0.206±0.003	0.369±0.013	0.528±0.024	0.702±0.015
T-CLIP (Global)	0.105±0.007	0.348±0.002	0.492±0.005	0.712±0.001	0.839±0.008	0.925±0.004
T-CLIP (F-G)	0.020±0.004	0.069±0.004	0.130±0.009	0.239±0.012	0.361±0.013	0.523±0.025
T-CLIP (Dual)	**0.117±0.003**	**0.364±0.008**	**0.511±0.005**	**0.728±0.005**	**0.851±0.006**	**0.932±0.005**
Δ vs. Global LoRA	+10.4%	+13.4%	+11.3%	+8.2%	+4.8%	+2.8%

Table 18: Text-image retrieval recall@K on the FMB (Liu et al., 2023) test set. All values reported as mean ± std across three independent training runs. **Bold**: best result. Underline: second best. F-G = fine-grained. Δ: relative R@K gain of the best T-CLIP variant per metric over Global LoRA.

IMAGE-TO-TEXT RETRIEVAL						
METHOD	**R@1**	**R@5**	**R@10**	**R@25**	**R@50**	**R@100**
CLIP (Global)	0.043	0.118	0.175	0.318	0.443	0.632
CLIP (F-G)	0.011	0.082	0.143	0.225	0.279	0.532
CLIP-Adapter (Global)	0.075±0.007	0.266±0.005	0.409±0.009	0.605±0.002	0.768±0.000	0.905±0.002
CLIP-Adapter (F-G)	0.029±0.004	0.157±0.011	0.277±0.005	0.455±0.016	0.677±0.002	0.880±0.009
DeCLIP (Global)	0.095±0.013	0.268±0.004	0.400±0.007	0.573±0.002	0.704±0.004	0.834±0.009
DeCLIP (F-G)	0.041±0.009	0.145±0.009	0.238±0.002	0.402±0.005	0.561±0.007	0.743±0.018
Global LoRA	0.104±0.015	0.342±0.012	0.496±0.016	0.731±0.019	0.866±0.012	0.956±0.010
F-G LoRA	0.037±0.002	0.133±0.014	0.248±0.015	0.479±0.038	0.658±0.032	0.842±0.015
T-CLIP (Global)	**0.105±0.012**	**0.362±0.007**	**0.539±0.018**	0.762±0.005	0.871±0.013	0.958±0.002
T-CLIP (F-G)	0.042±0.002	0.151±0.014	0.245±0.002	0.413±0.009	0.580±0.016	0.788±0.012
T-CLIP (Dual)	0.098±0.017	0.342±0.012	0.537±0.014	**0.771±0.008**	**0.895±0.007**	**0.973±0.005**
Δ vs. Global LoRA	+1.0%	+5.8%	+8.7%	+5.5%	+3.3%	+1.8%

TEXT-TO-IMAGE RETRIEVAL						
METHOD	**R@1**	**R@5**	**R@10**	**R@25**	**R@50**	**R@100**
CLIP (Global)	0.018	0.071	0.143	0.243	0.418	0.618
CLIP (F-G)	0.004	0.032	0.068	0.157	0.289	0.479
CLIP-Adapter (Global)	0.063±0.002	0.259±0.009	0.402±0.009	0.613±0.009	0.775±0.018	0.927±0.013
CLIP-Adapter (F-G)	0.039±0.004	0.148±0.009	0.243±0.004	0.438±0.013	0.611±0.014	0.863±0.016
DeCLIP (Global)	0.077±0.005	0.277±0.002	0.370±0.016	0.586±0.004	0.700±0.014	0.843±0.011
DeCLIP (F-G)	0.038±0.013	0.136±0.014	0.239±0.004	0.400±0.007	0.566±0.020	0.754±0.011
Global LoRA	0.100±0.012	0.350±0.024	0.510±0.020	0.737±0.016	0.882±0.018	0.960±0.010
F-G LoRA	0.048±0.002	0.185±0.018	0.282±0.008	0.489±0.005	0.666±0.019	0.848±0.036
T-CLIP (Global)	0.101±0.006	0.385±0.012	0.556±0.016	0.787±0.014	0.901±0.005	0.970±0.006
T-CLIP (F-G)	0.039±0.006	0.144±0.002	0.219±0.013	0.346±0.020	0.526±0.022	0.708±0.006
T-CLIP (Dual)	**0.104±0.015**	**0.395±0.034**	**0.587±0.009**	**0.818±0.016**	**0.914±0.003**	**0.979±0.008**
Δ vs. Global LoRA	+4.0%	+12.9%	+15.1%	+11.0%	+3.6%	+2.0%

Table 19: Full D2 physics correctness checklist (18 items, 7 scene-content groups). Fundamental rule: warm → brighter; cool → darker. Score = proportion YES among applicable items, mapped to 1–4 via Eq. 9. Typical applicable items per image: 6–9.

ITEM	NAME	WHAT TO CHECK	YES / NO CRITERION
Group A — Always applicable (no N/A)			
A1	Warm→brighter; Cool→darker	Both directions hold: warm bright AND cool dark	YES: bidirectional contrastNO: uniform or inverted
A2	Sky as cold thermal sink	Open sky one of darkest regions	YES: sky dark NO: sky warm
A3	Three-level hierarchy	Sky < Road < People/Vehicles	YES: three distinct levelsNO: road = sky or road = people
Group B — People present (N/A if no people)			
B1	Body heat distribution	Torso/head brightest; extremities darker	YES: upper body brighter than limbsNO: uniform glow
B2	Clothing insulation	Clothed areas slightly darker than bare skin	YES: clothed areas dimmerNO: uniform regardless of clothing
B3	Umbrella as cold shield	Umbrella dome dark — blocks body heat	YES: umbrella darker than personN/A: no umbrella
Group C — Vehicles present (N/A if no vehicles)			
C1	Engine/exhaust heat	Engine warm → brighter than body panels	YES: engine brighter NO: uniform
C2	Tyre heat	Tyres warm from friction → brighter than road	YES: tyres brighter than road
C3	Moving vs stationary	Moving warmer → brighter	YES: moving clearly brighterN/A: only one type present
C4	Headlights as bright spots	Headlights → brightest concentrated spots	YES: headlights brightestN/A: no night scene
Group D — Road visible (N/A if road not visible)			
D1	Road thermal state	Day: warm → bright; Night: intermediate	YES: correct for time of dayNO: road as dark as sky
D2	Wet surface emissivity	Wet ($\varepsilon \approx 0.97$) brighter than dry ($\varepsilon \approx 0.90$)	YES: wet areas brighter N/A: no wet surfaces
Group E — Adverse weather (N/A if clear)			
E1	Weather reduces contrast	Fog/rain → reduced warm/cool differentiation	YES: reduced contrast N/A: clear weather
E2	Overcast sky still dark	Overcast: sky dark but less extreme than clear night	YES: sky dark but not fully blackN/A: not overcast
Group F — Vegetation present (N/A if no vegetation)			
F1	Vegetation cooling	Evapotranspiration → vegetation darker than buildings	YES: vegetation clearly darkerNO: same as warm objects
Group G — Buildings present (N/A if no buildings)			
G1	Window-wall differential	Windows warm → brighter; walls cooler → darker	YES: windows brighter than wallNO: building uniformly bright
G2	Rooftop cooler than facade	Rooftop → darker than facade below	YES: rooftop darker N/A: rooftop not visible

Score mapping: 0–25% *YES → 1;* 26–50% *→ 2;* 51–75% *→ 3;* 76–100% *→ 4. If all N/A: exclude from D2 analysis.*

Table 20: Shared training configuration across all methods.

PARAMETER	VALUE
Base model	CLIP ViT-Base-Patch16
Input data	Thermal image–caption pairs (IR-Cap)
Optimizer	AdamW
β_1, β_2	0.9, 0.999
ϵ	1.0×10^{-8}
LR scheduler	Cosine with 100 warmup steps
Weight decay	1.0×10^{-3}
Max gradient norm	1.0
Train batch size	128
Gradient accumulation	1
Precision	FP16
Data workers	16
GPU	NVIDIA A6000 (48 GB)

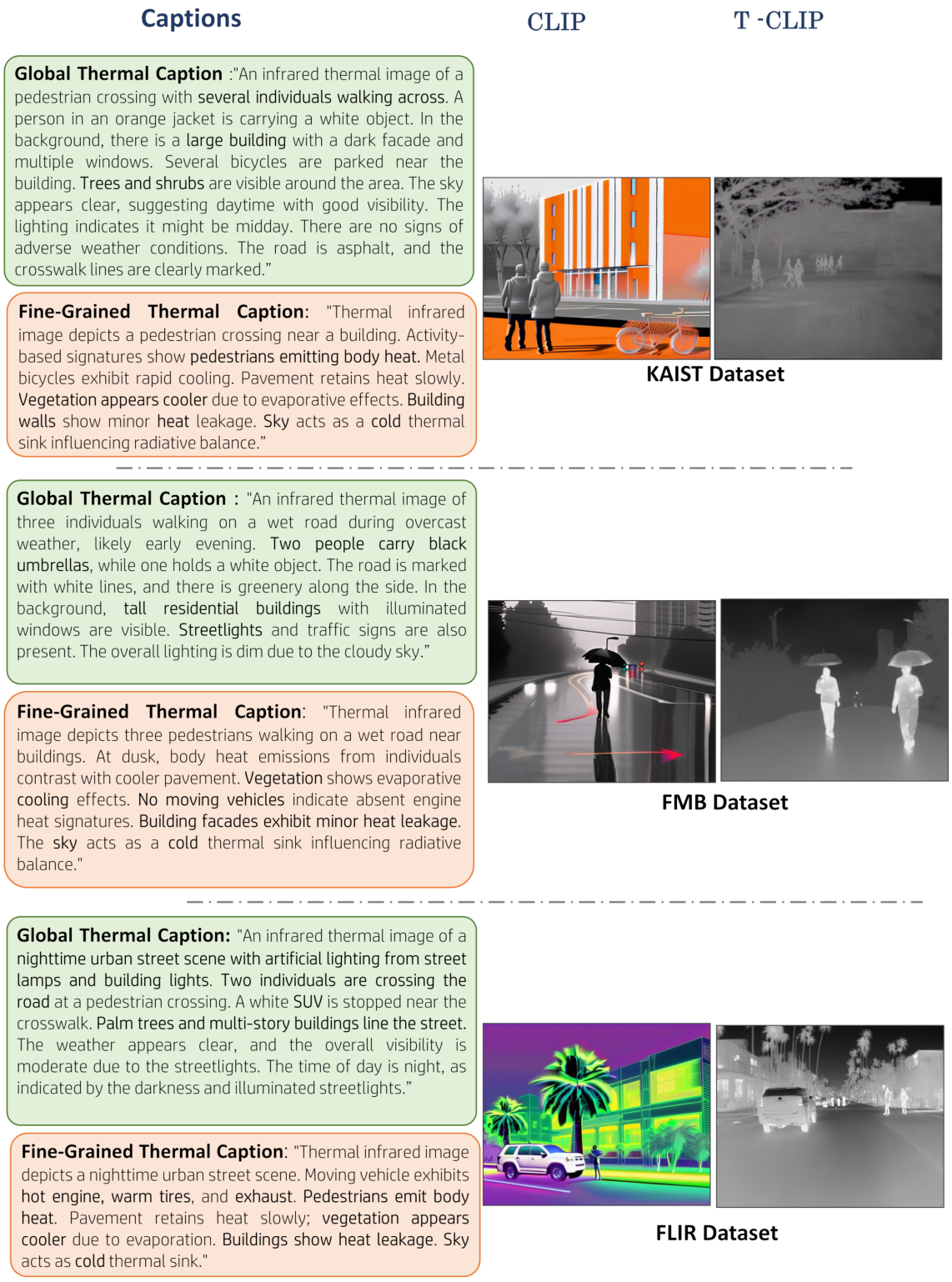

Figure 11: Generated thermal image samples using captions from KAIST (Hwang et al., 2015), FLIR (FLIR Systems, 2025), and FMB (Liu et al., 2023). For each prompt, left: zero-shot SDXL (standard CLIP text encoder); right: T-CLIP + SDXL. The T-CLIP + SDXL model captures both global scene context and fine-grained object-level heat signatures.

Figure 12: T-CLIP + SDXL performance under challenging conditions from FMB (Liu et al., 2023) dataset. For each pair, left: zero-shot SDXL (standard CLIP text encoder); right: T-CLIP + SDXL. T-CLIP + SDXL demonstrates thermal understanding in scenarios where conventional RGB-based approaches fail.

```
# V_g, V_f  –  Global / Fine-Grained LoRA Vision Encoder
# E_g, E_f  –  Global / Fine-Grained LoRA Text Encoder
# I         –  minibatch of thermal images
# C_g       –  minibatch of global thermal captions
# C_fg      –  minibatch of fine-grained thermal captions
# alpha     –  ensemble weight  (default: 0.8)
```

Training

```
# Global LoRA branch
img_g   =  V_g( I )
txt_g   =  E_g( C_g )
loss_g  =  clip_loss( img_g, txt_g, tau )

# Fine-Grained LoRA branch
img_f   =  V_f( I )
txt_f   =  E_f( C_fg )
loss_f  =  clip_loss( img_f, txt_f, tau )

# Branches updated independently – no gradient interference
loss_g.backward()                    # updates theta_g only
loss_f.backward()                    # updates theta_f only
```

Inference

```
# Step 1: Independent branch encoding
img_g   =  L2_norm( V_g( I ) )
txt_g   =  L2_norm( E_g( C_g ) )
img_f   =  L2_norm( V_f( I ) )
txt_f   =  L2_norm( E_f( C_fg ) )

# Step 2: Weighted fusion in embedding space
img_fused  =  alpha * img_g  +  (1-alpha) * img_f
txt_fused  =  alpha * txt_g  +  (1-alpha) * txt_f

# Step 3: L2 normalize after fusion
img_fused  =  L2_norm( img_fused )
txt_fused  =  L2_norm( txt_fused )

# Step 4: Retrieval similarity
s  =  dot( img_fused,  txt_fused.T )
```

Figure 13: T-CLIP training and inference pseudocode. The global LoRA (θ_g) and fine-grained LoRA (θ_f) branches are optimised independently on semantically distinct caption types with no shared gradient flow, preventing representational interference. During inference, L2-normalised features from both branches are fused via the ensemble weight α=0.8 before final similarity computation (see section 4.4).

Table 21: Method-specific hyperparameters. LoRA (Global) and LoRA (F-G) share identical settings, trained on global and fine-grained captions respectively. DeCLIP uses a lower learning rate. "—" denotes not applicable.

PARAMETER	**LORA (GLOBAL / F-G)**	**CLIP-ADAPTER (Gao et al., 2024)**	**DECLIP (Li et al., 2022b)**
Trainable parameters	LoRA adapters only	Bottleneck MLP only	All CLIP parameters
Frozen backbone	✓	✓	✗
LoRA rank r	8	—	—
LoRA α	1	—	—
LoRA dropout	0	—	—
Target modules	`q,k,v_proj`	—	—
Adapter blend ratio	—	0.2	—
Adapter bottleneck	—	512→128→512	—
Learning rate	2.0×10^{-3}	1.0×10^{-3}	1.0×10^{-5}
Max steps	10,000	10,000	10,000
Training time	≈1.5 hrs	≈40 min	≈2 hrs
λ_{ii} (image-image weight)	—	—	0.5
λ_{tt} (text-text weight)	—	—	0.5
NN temperature τ	—	—	0.1
Top-k NNs	—	—	1
Momentum m (EMA)	—	—	0.995
Memory bank size	—	—	4,096

LoRA target modules: `q_proj`, `k_proj`, `v_proj` in both vision and text encoders. CLIP-Adapter blend: output $= 0.2 \times \text{adapter}(x) + 0.8 \times x$. DeCLIP memory bank stores the last 4,096 momentum encoder features in a FIFO queue; NNs searched over the combined bank and current batch for stable target computation.