FROM UNPAIRED PRIORS TO PAIRED CALIBRATION: T2IR FOR RGB-TO-INFRARED IMAGE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB-to-infrared (IR) image translation is difficult because visible reflected light and thermal radiation follow fundamentally different physical processes. Paired RGB/IR data are also scarce and frequently only weakly aligned, so a single-paradigm approach often fails to jointly maintain structural correspondence and IR appearance diversity. To tackle these issues, we propose T2IR, a two-stage diffusion framework that moves from unpaired prior learning to paired structural calibration. The first stage exploits separate RGB and IR image pools to acquire a cross-modal infrared appearance prior. The second stage uses paired supervision to adjust geometric structure and infrared intensity. We additionally apply training-time regularization on thermal sources and texture statistics, which mitigates underheated thermal regions and visible-texture leakage from RGB images. To support this framework, we construct THIRAlign, a synchronized paired RGB/IR dataset obtained with a beam-splitter-assisted dual-modal acquisition system and calibration-based spatial alignment. Experiments on three datasets demonstrate that T2IR surpasses six state-of-the-art methods. The results verify that unpaired infrared prior learning and paired structural calibration are complementary, and that the proposed training-time regularization compensates for the weakness of global image reconstruction losses.

1 INTRODUCTION

Many visual perception models are designed and trained primarily using visible-light images captured under favorable illumination conditions. In nighttime, low-light, adverse-weather, or otherwise complex environments, visible-light images often suffer from insufficient brightness, reduced contrast, and degraded texture, which can degrade downstream performance (Hwang et al., 2015; Jia et al., 2021). Unlike visible-light imaging, thermal infrared sensors, especially long-wave infrared (LWIR) sensors, measure emitted and reflected thermal radiation (Vollmer & Möllmann, 2017; Bao et al., 2023) and are less dependent on external visible-light illumination (Vollmer & Möllmann, 2017). They can therefore provide complementary target information in nighttime and low-light environments (Jia et al., 2021), supporting applications such as aerial surveillance and low-light detection (Sun et al., 2022; Zhang et al., 2022), as well as cross-modal data augmentation (Ran et al., 2025).

Converting readily available visible-light images into thermal infrared images offers a promising technical approach to expanding infrared datasets and supporting cross-modal visual perception (Xiao et al., 2025; Ran et al., 2025). However, visible-light and thermal infrared images are produced by fundamentally different imaging mechanisms. Visible-light cameras primarily record ambient-light reflection from objects, representing color, brightness, and surface texture. Thermal infrared sensors, by contrast, respond to target temperature and material emissivity (Vollmer & Möllmann, 2017; Usamentiaga et al., 2014), as well as reflected environmental radiation, wavelength band, and sensor response (Bao et al., 2023; Mao et al., 2026). Without information about these acquisition and environmental factors, the RGB-to-infrared mapping is underdetermined rather than a one-to-one correspondence. Similar visible textures can therefore correspond to different thermal-radiation levels (Bao et al., 2023), while prominent heat sources can lack distinctive visible-light features (Mao et al., 2026). An effective RGB-to-infrared generation model must not only preserve the spatial layout, object locations, and geometric structure of the input scene, but also learn infrared appearance priors and intensity distributions that are consistent with thermal imaging principles.

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



Figure 1: Examples from THIRAlign. The top row shows visible-light (RGB) images and the bottom row shows their paired infrared (IR) observations. The collection includes ground-level road scenes and elevated road views captured from high-rise buildings.

Although RGB-to-infrared image translation has made progress in recent years, the task remains constrained by both data acquisition and thermal imaging modeling. On the one hand, collecting high-quality thermal infrared data requires dedicated sensors and complex imaging systems, while supervised training generally requires RGB–IR image pairs that are aligned in both space and time (Xiao et al., 2025; Ma et al., 2024). Such data require synchronized dual-modal sensing and calibration to mitigate sensor parallax, temporal asynchrony, object motion, field-of-view differences, and cross-modal registration errors (Jia et al., 2021; Hwang et al., 2015). Despite a growing number of public paired datasets, their coverage is often constrained by sensing setup, viewpoint, scene composition, and acquisition conditions, which can limit cross-application generalization (Han et al., 2023; Xiao et al., 2025).

On the other hand, early RGB-to-infrared methods commonly formulate the task as appearance or style translation for person re-identification (Kniaz et al., 2018), generic scenes (Özkanoğlu & Ozer, 2022), and aerial imagery (Ma et al., 2024). More recent work introduces semantic priors (Ran et al., 2025) or physics-informed priors (Mao et al., 2026). Nevertheless, thermal fidelity remains challenging: existing studies report retained visible-light artifacts and visible-texture leakage (Xiao et al., 2025; Ran et al., 2025), physically incorrect relative thermal responses, and errors around bright sources or local high-intensity regions (Mao et al., 2026). The central difficulty is therefore a mismatch between the available supervision regimes. Large unpaired RGB and IR collections provide broad domain coverage but lack scene-wise correspondence, whereas paired data offer explicit cross-modal correspondence but remain limited and may contain residual misalignment. Moreover, conventional global reconstruction objectives do not explicitly constrain local thermal characteristics, such as compact heat-source responses, bright-tail intensity statistics, and the suppression of RGB-only texture leakage.

To address these issues, we propose *T2IR*, a two-stage diffusion framework that combines unpaired prior learning with paired thermal calibration. Stage 1 learns a bidirectional RGB–IR appearance prior from unpaired image pools, while Stage 2 initializes the RGB$\rightarrow$IR path from this prior and fine-tunes it using paired RGB–IR supervision. We further introduce training-only thermal-appearance regularization to improve bright-tail and compact-source responses and suppress RGB-only texture leakage.

The main contributions of this paper are summarized as follows:

- **THIRAlign Dataset.** We construct a synchronized and calibration-aligned RGB–IR dataset with sequence-level splits across ground-level and elevated road scenes. It provides reliable paired data for training and controlled evaluation.
- **Two-Stage Diffusion Framework.** We combine unpaired RGB–IR prior learning with paired RGB-to-IR calibration in a sequential training framework. Experiments show that the two supervision regimes provide complementary improvements over either stage alone.
- **Thermal-Appearance Regularization.** We introduce training-only statistical constraints for bright-tail responses, compact thermal sources, thermal fields, and RGB-only edge leakage. The corresponding ablations demonstrate improved thermal fidelity and reduced visible-texture transfer.

2 Related Work

RGB–IR datasets and alignment. Paired visible–infrared datasets support multispectral perception across ground-level, aerial, and low-light scenarios: KAIST and LLVIP focus on pedestrian-centric scenes, M3FD and MSRS cover broader environments, and VTUAV and AVIID extend collection to aerial platforms (Hwang et al., 2015; Jia et al., 2021; Liu et al., 2022; Tang et al., 2022; Zhang et al., 2022; Han et al., 2023). Differences in sensing setups and acquisition conditions, together with parallax, temporal offsets, object motion, and registration errors, can leave residual misalignment and limit direct pixel-wise supervision (Wang et al., 2022; Ordun et al., 2023). THIRAlign complements these datasets with synchronized beam-splitter-assisted acquisition and calibration-based alignment across ground-level and elevated road scenes.

RGB-to-infrared translation. Early RGB-to-infrared methods use adversarial image-to-image translation for person-oriented, generic-scene, and aerial generation (Kniaz et al., 2018; Özkanoğlu & Ozer, 2022; Ma et al., 2024). Recent work adopts diffusion or flow-based models, while semantic guidance, foundation-model priors, and physics-inspired decomposition further improve cross-modal modeling and infrared appearance (Huang et al., 2023; Hu et al., 2025; Xiao et al., 2025; Ran et al., 2025; Paranjape et al., 2026; Mao et al., 2026). Because RGB appearance does not uniquely determine thermal response, existing methods may retain visible textures, underestimate compact high-intensity regions, or produce implausible intensity distributions; T2IR addresses these issues through unpaired infrared-prior learning, paired calibration, and training-time thermal-appearance regularization.

Mixed supervision and thermal-aware regularization. Paired translation methods such as pix2pix learn explicit source–target correspondence, whereas unpaired methods such as CycleGAN and CUT exploit larger domain-specific collections without identifying the response of a particular RGB scene (Isola et al., 2017; Zhu et al., 2017; Park et al., 2020). Related paired and unpaired formulations have also been studied with diffusion bridges and neural Schrödinger bridges (Li et al., 2023; Liu et al., 2023; Kim et al., 2024). T2IR combines the two regimes sequentially, using shift-tolerant paired supervision and statistical constraints on bright-tail responses, compact thermal sources, thermal fields, and RGB-only edge leakage.

3 Method

3.1 Preliminaries

T2IR instantiates its cross-modal generator with a latent-diffusion translator (Rombach et al., 2022; Parmar et al., 2024). For source and target domains $a, b \in \{\mathrm{RGB}, \mathrm{IR}\}$, respectively, an input u from domain a is translated toward domain b as

$$\begin{aligned} z &= \mathrm{Enc}_a(u), \\ \hat{z} &= S(z, U_\theta(z, t, \tau(p_b)), t)\,, \\ F_{a\to b}(u) &= \mathrm{Dec}_b(\hat{z}). \end{aligned} \tag{1}$$

Here, Enc_a and Dec_b are domain-specific VAE encoders and decoders (Rombach et al., 2022), τ is the CLIP text encoder (Radford et al., 2021), U_θ is a shared conditional UNet (Rombach et al., 2022), and S denotes the scheduler update at the fixed diffusion timestep t (Parmar et al., 2024). The two directions share U_θ. We use the fixed target-domain prompts $p_{\mathrm{RGB}} =$ `visible image` and $p_{\mathrm{IR}} =$ `thermal infrared image`, with p_b denoting the prompt for target domain b. Following the translator, the generator uses trainable 1×1 encoder–decoder skip convolutions (Parmar et al., 2024) to preserve low-level spatial detail.

3.2 Overview of T2IR

RGB-to-IR translation presents a supervision mismatch: large unpaired image pools provide a broad infrared appearance prior but no scene-wise correspondence, whereas paired data provide correspondence but are scarce and weakly aligned (Zhu et al., 2017; Isola et al., 2017; Hwang et al., 2015; Jia et al., 2021). T2IR resolves this mismatch through sequential prior transfer and paired thermal

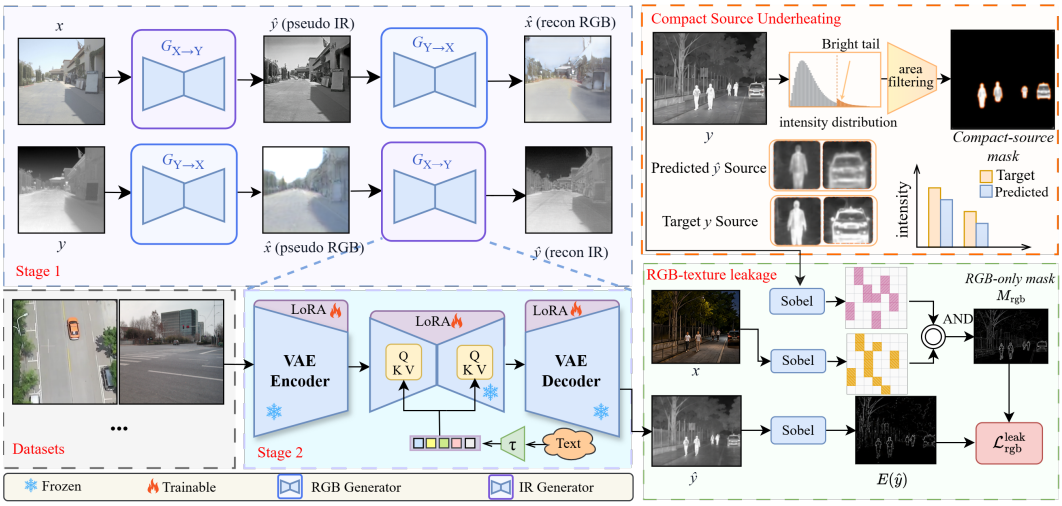

Figure 2: T2IR framework. Stage 1 learns an infrared appearance prior through bidirectional translation using large-scale unpaired RGB and IR data, including IR-500K (Ran et al., 2025). Stage 2 fine-tunes the RGB-to-IR path on multiple public datasets with a paired objective and a training-only thermal-statistics regularizer. All test samples and their duplicates or counterparts are excluded from training.

calibration. It first learns an RGB–IR prior from unpaired images and then selectively calibrates the active RGB→IR path with paired supervision.

Let $x \in \mathcal{X}$ be an RGB image and $y \in \mathcal{Y}$ its approximately aligned IR observation. We denote paired samples by $\mathcal{D}_p = \{(x_i, y_i)\}$ and modality-specific image pools by $\mathcal{D}_u = (\{x_i\}, \{y_j\})$. We write $\hat{y} = F_{\mathrm{RGB}\to\mathrm{IR}}(x)$ for the generated infrared image.

Stage 1: unpaired cross-modal prior learning. Stage 1 treats the samples in $\mathcal{D}_u$ as independent images and trains the bidirectional backbone with cycle, identity, and adversarial objectives (Zhu et al., 2017; Parmar et al., 2024). This establishes a transferable RGB–IR mapping prior without using pair identities. The resulting Stage-1 generator parameter set $\Theta_G^{(1)}$ contains the trainable UNet input convolution and LoRA adapters (Hu et al., 2022), VAE adapters, and decoder skip convolutions. Since its samples are unpaired, this checkpoint captures domain-level infrared appearance but cannot specify the thermal response of an individual RGB scene.

Stage 2: paired thermal-response calibration. Rather than training an independent paired model, we transfer the RGB→IR-relevant tensors from $\Theta_G^{(1)}$ and fine-tune only their effective forward path: the UNet input convolution and LoRA adapters, VAE encoder adapters, VAE decoder adapters and skip convolutions. This selective update inherits and selectively refines the bidirectionally learned domain prior while allocating scarce paired supervision to the parameters that determine RGB-to-IR responses. Formally, let $\Theta_{\mathrm{path}}^{(1)} \subset \Theta_G^{(1)}$ denote the Stage-1 tensors on this active path. We initialize $\Theta_{\mathrm{path}} \leftarrow \Theta_{\mathrm{path}}^{(1)}$ before paired optimization and obtain

$$\Theta_{\mathrm{path}}^{(2)} = \arg\min_{\Theta_{\mathrm{path}}} \mathcal{L}_{\mathrm{stage2}}(F_{\mathrm{RGB}\to\mathrm{IR}}(x), y, x). \tag{2}$$

Only $F_{\mathrm{RGB}\to\mathrm{IR}}$ is retained at inference; the reverse translator, discriminators, perceptual network, and statistical masks are used only during training, so T2IR adds no inference-time branch or sampling step.

3.3 Two-Stage Optimization

Stage-1 prior learning. Stage 1 optimizes the adopted bidirectional backbone with

$$\mathcal{L}_{\mathrm{stage1}} = \mathcal{L}_{\mathrm{cyc}} + \mathcal{L}_{\mathrm{idt}} + \lambda_{\mathrm{gan}}\mathcal{L}_{\mathrm{gan}}. \tag{3}$$

where $\mathcal{L}_{\text{cyc}}$, $\mathcal{L}_{\text{idt}}$, and $\mathcal{L}_{\text{gan}}$ denote the cycle-consistency, identity, and adversarial losses, respectively, and λ_{gan} controls the contribution of the adversarial term.

Stage-2 objective. Stage 2 deactivates the reverse translation path and discards the discriminators. Starting from $\Theta_{\text{path}} \leftarrow \Theta^{(1)}_{\text{path}}$, it optimizes the forward-path parameters with a paired objective, yielding $\Theta^{(2)}_{\text{path}}$. We separate reconstruction terms, alignment-and-leakage terms, and source-and-field thermal-statistics terms.

$$\mathcal{L}_{\text{stage2}} = \boldsymbol{\lambda}_r^\top \boldsymbol{\ell}_r + \boldsymbol{\lambda}_{st}^\top \boldsymbol{\ell}_{st} + \boldsymbol{\lambda}_t^\top \boldsymbol{\ell}_t, \tag{4}$$

where $\boldsymbol{\ell}_r$, $\boldsymbol{\ell}_{st}$, and $\boldsymbol{\ell}_t$ collect the reconstruction terms, alignment-and-leakage terms, and source-and-field thermal-statistics terms, respectively. For the full Stage-2 objective, we use $\boldsymbol{\lambda}_r = (1.0, 1.5)^\top$, $\boldsymbol{\lambda}_{st} = (2.0, 0.10)^\top$, and $\boldsymbol{\lambda}_t = (0.5, 0.5, 1.0, 0.75)^\top$.

Reconstruction and alignment-and-leakage terms. The reconstruction and alignment-and-leakage terms are collected as

$$\begin{aligned} \boldsymbol{\ell}_r &= \left[\mathcal{L}_1(\hat{y}, y), \mathcal{L}_{\text{LPIPS}}(\hat{y}, y)\right]^\top, \\ \boldsymbol{\ell}_{st} &= \left[\mathcal{L}_{\text{si}}, \mathcal{L}^{\text{leak}}_{\text{RGB}}\right]^\top. \end{aligned} \tag{5}$$

Here $\boldsymbol{\lambda}_r = (\lambda_1, \lambda_p)^\top$ and $\boldsymbol{\lambda}_{st} = (\lambda_{\text{si}}, \lambda_{\text{leak}})^\top$. The alignment-and-leakage terms enforce shift-tolerant correspondence and suppress RGB-only texture transfer. To tolerate small sensor offsets observed in paired visible–infrared capture (Hwang et al., 2015; Jia et al., 2021) and addressed by misalignment or registration methods (Wang et al., 2022; Ordun et al., 2023), let $\delta = (\delta_u, \delta_v)$ denote horizontal and vertical integer displacements, respectively, and let P_δ and Y_δ be the corresponding candidate-dependent valid overlapping crops of $\hat{y}$ and y. We use

$$\begin{aligned} \delta^* &= \arg \min_{\delta_u, \delta_v \in \{-r, \ldots, r\}} \text{MSE}(P_\delta, Y_\delta), \\ \mathcal{L}_{\text{si}} &= \text{MSE}(P_{\delta^*}, Y_{\delta^*}). \end{aligned} \tag{6}$$

With $r = 2$, the optimal displacement is selected independently for each image.

For RGB-texture leakage suppression, let $I(z)$ be the channel-averaged intensity map of z after rescaling to $[0, 1]$, and let $E(z)$ be its Sobel magnitude. Denote by $Q_\eta(E(z))$ the empirical η-quantile of the spatial values in $E(z)$, where $\eta \in (0, 1)$, and set $\tau_z = Q_\eta(E(z))$. Using the indicator $\mathbb{1}[\cdot]$, which is one when its condition holds and zero otherwise, the RGB-only mask is

$$M_{\text{RGB}} = \mathbb{1}[E(x) \geq \tau_x] \odot \mathbb{1}[E(y) < \tau_y]. \tag{7}$$

Here $\odot$ denotes element-wise multiplication. For a binary spatial mask M and a scalar-valued spatial map a, where p indexes spatial positions and a_p is the value of a at p, define the masked ℓ_1 mean as

$$\|a\|_{1,M} = \frac{\sum_p M(p)|a_p|}{\max(1, \sum_p M(p))}. \tag{8}$$

Setting $a = E(\hat{y})$ and $M = M_{\text{RGB}}$ gives the leakage penalty

$$\mathcal{L}^{\text{leak}}_{\text{RGB}} = \|E(\hat{y})\|_{1, M_{\text{RGB}}}, \tag{9}$$

which penalizes generated IR edges where RGB has a strong edge but the observed IR image lacks a corresponding strong edge, thereby suppressing RGB-only texture transfer.

Source-and-field thermal-statistics terms. To capture thermal-image characteristics that are insufficiently constrained by the reconstruction and alignment-and-leakage terms, the source-and-field thermal-statistics vector contains source-response and thermal-field components motivated by temperature and emissivity factors (Vollmer & Möllmann, 2017; Usamentiaga et al., 2014), radiometric measurements (Bao et al., 2023), and physics-informed infrared generation (Mao et al., 2026):

$$\boldsymbol{\ell}_t = \left[\mathcal{L}_{\text{quantile}}, \mathcal{L}_{\text{cc}}, \mathcal{L}_{lf}, \mathcal{L}_{hf}\right]^\top. \tag{10}$$

Here $\boldsymbol{\lambda}_t = (\lambda_q, \lambda_{\text{cc}}, \lambda_{lf}, \lambda_{hf})^\top$.

Thermal source response. The source-response terms address global bright-tail mismatch and local under-heating of compact sources. Let $Q_q(\cdot)$ denote the empirical per-image q-quantile, with $\mathcal{Q} = \{0.85, 0.90, 0.95, 0.98\}$. Let M_c be a binary mask obtained from upper-tail connected components of $I(y)$ using fixed area and shape criteria, and let p index spatial positions. We define

$$\mathcal{L}_{\text{quantile}} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \left| Q_q(I(\hat{y})) - Q_q(I(y)) \right|, \tag{11}$$

$$\mathcal{L}_{\text{cc}} = \frac{\sum_p M_c(p) \left[I(y)_p - I(\hat{y})_p \right]_+^\gamma}{\max(1, \sum_p M_c(p))}, \tag{12}$$

where $[t]_+ = \max(t, 0)$, $\gamma = 1.5$, and the denominator prevents division by zero when M_c is empty. The quantile term matches the global bright-tail distribution, while $\mathcal{L}_{\text{cc}}$ penalizes only underestimated intensities within compact thermal-source regions, with $\gamma > 1$ emphasizing larger under-heating errors.

Thermal field. The thermal-field terms separate coarse thermal layout from high-frequency structure. Let G_σ be a normalized Gaussian kernel with standard deviation σ, $B(z) = G_\sigma * I(z)$ the resulting low-pass field, and $H(z) = I(z) - B(z)$ its high-frequency residual. For an unmasked residual, $\| \cdot \|_1$ denotes the mean absolute value over spatial positions. The common-edge mask is

$$M_{ce} = \mathbb{1}[E(x) \geq \tau_x] \odot \mathbb{1}[E(y) \geq \tau_y]. \tag{13}$$

The thermal-field terms are

$$\mathcal{L}_{lf} = \| B(\hat{y}) - B(y) \|_1, \tag{14}$$

$$\mathcal{L}_{hf} = \| H(\hat{y}) - H(y) \|_{1, M_{ce}}. \tag{15}$$

These terms match the low-frequency thermal field and high-frequency residuals at cross-modal common-edge locations, respectively. They and the masks defined above are applied only during training. The remaining implementation details and fixed parameter settings are specified in the supplementary technical appendix A.

4 Experiments

4.1 Settings

Dataset. We construct THIRAlign using a beam-splitter-assisted dual-modal imaging rig. The visible camera uses a Sony IMX385LQR-C CMOS sensor with a native resolution of 1920×1080, while the thermal camera uses an uncooled vanadium-oxide LWIR detector with a resolution of 640×512, an 8–14 μm response band, and an NETD of at most $40\,\text{mK}$. The cameras are mounted with approximately orthogonal physical axes around the beam splitter. Through its transmitted and reflected optical paths, their effective viewing directions are made approximately coaxial, thereby reducing cross-modal parallax.

Frames from the two cameras are associated using shared timestamps, and only RGB–IR pairs with an absolute timestamp difference of no more than $20\,\text{ms}$ are retained. RGB images are cropped and resampled to match the thermal camera's effective field of view and native 640×512 resolution. Cross-modal spatial alignment is performed using a calibration board containing features observable in both modalities. Corresponding points extracted from multiple calibration views are used to estimate a transformation that maps the infrared images into the visible-image coordinate system.

The resulting dataset contains 32,214 temporally matched and calibration-aligned RGB–IR pairs from ground-level and elevated road scenes. We divide the data into 28,887 training pairs and 3,327 test pairs by contiguous capture sequence, preventing temporally adjacent frames from crossing the split.

We manually construct the train/test partitions for all data sources before model training and hyperparameter tuning. The resulting partitions are fixed and shared across all evaluated methods. All evaluation samples, identifiable duplicates, and cross-modal counterparts are excluded when constructing the training pools. Stage 1 uses the retained training images from IR-500K (Ran et al.,

Table 1: Unified RGB-to-IR comparison on KAIST, LLVIP, and THIRAlign using PSNR, SSIM, and LPIPS. Best and second-best results are shown in bold and underlined, respectively.

Methods	Venue	KAIST			LLVIP			THIRAlign		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
CycleGAN-Turbo	arXiv 2024	13.51	0.322	0.655	11.07	0.201	0.701	10.80	0.189	0.559
StegoGAN	CVPR 2024	10.95	0.245	0.570	12.35	0.234	0.654	11.25	0.225	0.536
pix2pix-Turbo	arXiv 2024	25.98	0.779	0.382	15.36	0.451	0.463	17.46	0.424	0.319
PID	PR 2026	28.42	<u>0.843</u>	0.375	14.20	0.416	0.565	18.15	0.416	0.438
DiffV2IR	arXiv 2025	<u>29.49</u>	0.828	<u>0.359</u>	20.96	<u>0.701</u>	0.395	22.06	0.622	<u>0.264</u>
ThermalGen	NeurIPS 2025	29.38	0.797	0.407	<u>21.18</u>	0.671	<u>0.384</u>	<u>23.28</u>	<u>0.657</u>	0.265
T2IR	–	**31.56**	**0.898**	**0.327**	**21.76**	**0.714**	**0.354**	**24.18**	**0.692**	**0.243**

2025), THIRAlign, and the other available data sources as independently sampled RGB or IR images, without exploiting RGB–IR pair identities. Thus, images originating from paired datasets are included in Stage 1, but their cross-modal correspondence is discarded. Stage 2 uses only paired RGB–IR samples from the manually constructed training splits of THIRAlign, KAIST, and LLVIP (Hwang et al., 2015; Jia et al., 2021), M3FD (Liu et al., 2022), MSRS (Tang et al., 2022), VTUAV and AVIID (Zhang et al., 2022; Han et al., 2023), and DroneVehicle (Sun et al., 2022), and FLIR (Teledyne FLIR, 2018), explicitly exploiting their pair correspondence. No unpaired or visible-only images are used for Stage 2 supervision. Details of the dataset composition and split protocol are provided in the supplementary technical appendix B.

Training details. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU. For joint training across multiple datasets, each image is randomly resized and cropped to 256×256 pixels. During Stage 1, the RGB-to-IR branch uses the fixed target-domain prompt `thermal infrared image`, whereas the IR-to-RGB branch uses `visible image`. For Full T2IR, Stage 1 is optimized for $30{,}000$ steps with AdamW (Loshchilov & Hutter, 2019) at a learning rate of 10^{-5}, while Stage 2 is optimized for $20{,}000$ steps with AdamW at a learning rate of 5×10^{-6}. The remaining implementation details are reported in the supplementary technical appendix A.

Evaluation metrics. We evaluate generated infrared images against their paired ground truth using peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) (Wang et al., 2004), and learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018). Higher PSNR and SSIM indicate better reconstruction fidelity and structural similarity, respectively, whereas lower LPIPS indicates greater perceptual similarity.

4.2 Comparison with Other Methods

We reproduce and evaluate CycleGAN-Turbo (Parmar et al., 2024), StegoGAN (Wu et al., 2024), pix2pix-Turbo (Parmar et al., 2024), PID (Mao et al., 2026), DiffV2IR (Ran et al., 2025), and ThermalGen (Xiao et al., 2025) under a unified evaluation protocol. For training, CycleGAN-Turbo and StegoGAN, which support unpaired image translation, are trained using the same aggregated training data as T2IR. In contrast, methods that require paired supervision are trained on the union of all available paired RGB–thermal samples in the aggregated dataset. Thus, each baseline is trained with the largest applicable subset of the same data collection, subject to its original training requirements. All methods are reproduced using their official implementations and recommended training settings.

Table 1 shows that T2IR achieves the best PSNR, SSIM, and LPIPS on all three datasets. The consistent gains across fidelity and perceptual metrics indicate that the method improves both pixel-level

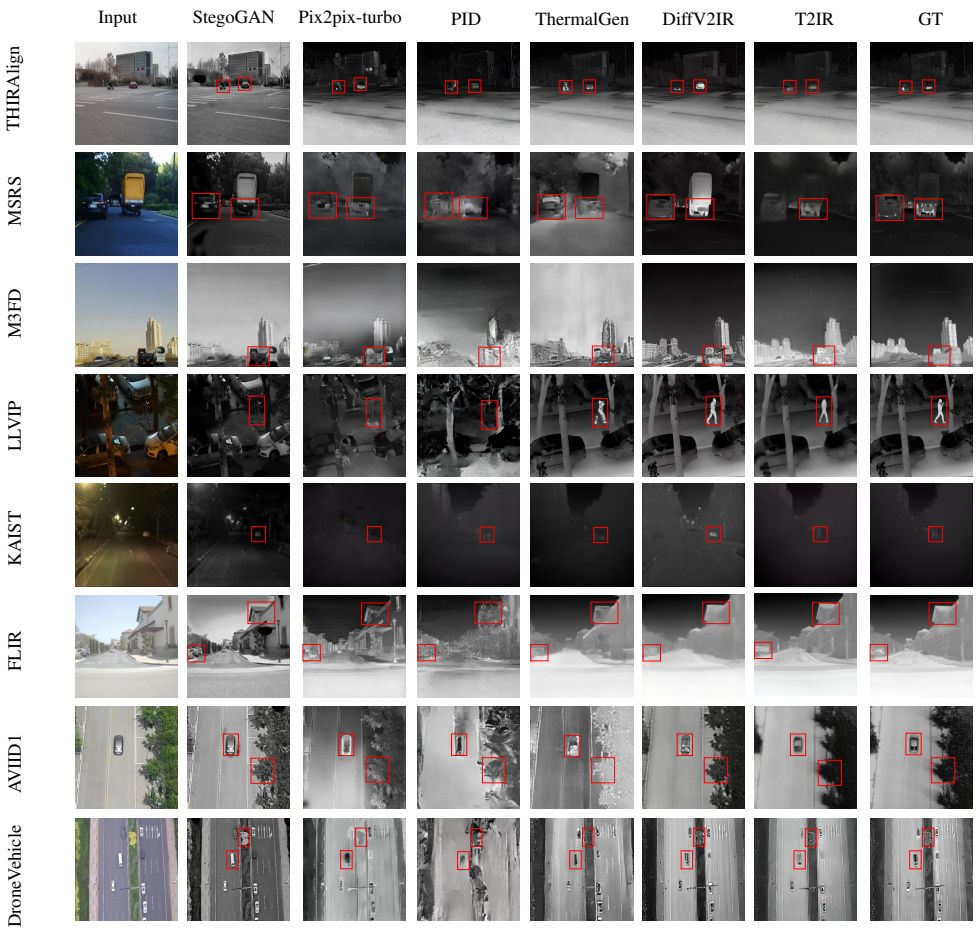

Figure 3: Qualitative RGB-to-IR comparison across eight data samples. Rows, from top to bottom, are THIRAlign, MSRS, M3FD, LLVIP, KAIST, FLIR, AVIID1, and DroneVehicle. Columns are Input, StegoGAN, Pix2pix-turbo, PID, ThermalGen, DiffV2IR, T2IR, and ground truth (GT).

reconstruction and cross-modal appearance quality across independently evaluated data distributions.

The visualization results in Fig. 3 demonstrate the superior performance of T2IR. Across the displayed road, surveillance, aerial, and low-light scenes, T2IR preserves the principal scene layout while producing infrared responses that are visually closer to the paired target. In particular, it better retains compact bright-source responses and suppresses visible-spectrum color and texture transfer in regions where the target infrared image is structurally smoother.

4.3 Stage and Loss Ablations

Table 2 compares Stage-1 Only, Stage-2 Only, a matched two-stage L1+LPIPS baseline, and Full T2IR on the held-out THIRAlign split. Stage 2 Only outperforms Stage 1 Only, while Full T2IR further improves over Stage 2 Only by 4.64 dB in PSNR and 0.189 in SSIM, and reduces LPIPS by 0.079.

Against the matched two-stage L1+LPIPS baseline, Full T2IR gains 1.60 dB in PSNR and 0.090 in SSIM, while reducing LPIPS by 0.053. These results support the benefit of the complete Stage 2 objective beyond the training schedule alone. Figure 4 provides complementary qualitative evidence: Full T2IR produces more coherent thermal fields and stronger compact-source responses while preserving scene structure.

Table 2: Stage and design ablation on the held-out THIRAlign test split using PSNR, SSIM, and LPIPS. U/P denote unpaired/paired training; the matched two-stage baseline uses L1+LPIPS.

Configuration	U	P	PSNR ↑	SSIM ↑	LPIPS ↓
Stage-1 Only	yes	no	14.32	0.423	0.433
Stage-2 Only	no	yes	19.54	0.503	0.322
two-stage L1+LPIPS	yes	yes	22.58	0.602	0.296
Full T2IR	yes	yes	**24.18**	**0.692**	**0.243**

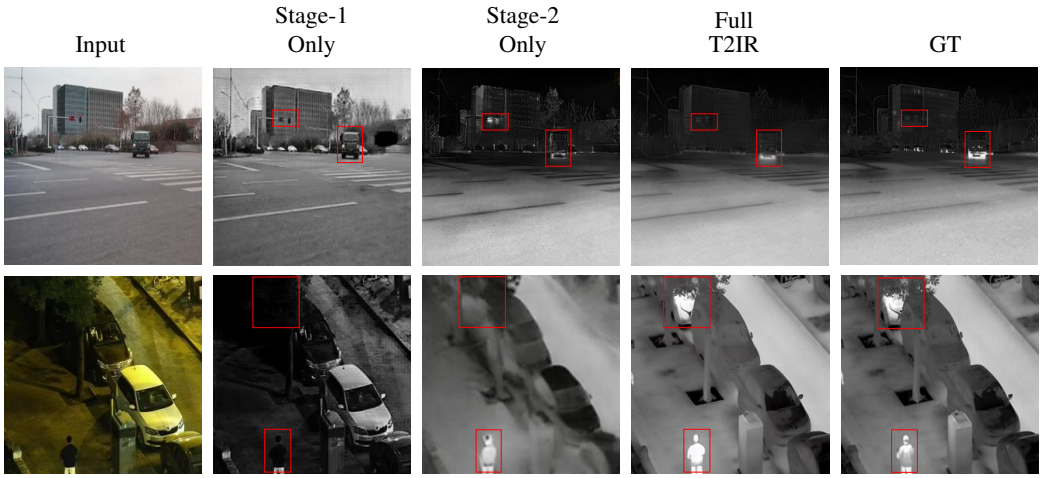

Figure 4: Visual comparison of Stage-1 Only, Stage-2 Only, and Full T2IR. The top and bottom rows show representative samples from LLVIP and THIRAlign, respectively. Columns are Input, Stage-1 Only, Stage-2 Only, Full T2IR, and ground truth (GT).

Additional controlled loss ablations are reported in Appendix C. Removing the perceptual, shift-tolerant, source-response, thermal-field, or leakage-suppression terms degrades all three reported metrics, supporting the use of the complete objective.

4.4 Limitations

T2IR incurs additional training cost because unpaired bidirectional pretraining precedes paired calibration. Stage 2 still requires approximately aligned RGB–IR pairs, and the shift-tolerant objective handles only small translations. The thermal-appearance losses also use fixed thresholds and filtering scales whose transfer across resolutions and intensity distributions remains to be studied. Future work will investigate efficient prior transfer, few-pair calibration, learned correspondence, adaptive multi-scale regularization, and downstream applications of generated infrared images.

5 Conclusion

We presented T2IR, a two-stage RGB-to-IR diffusion framework that combines unpaired cross-modal prior learning with paired structural and intensity calibration. Thermal-aware training objectives and the synchronized, calibration-aligned THIRAlign dataset further address weak alignment and visible-texture leakage. Under a unified evaluation protocol, T2IR improves over six competing methods on three held-out datasets, while controlled ablations show complementary benefits from the Stage-1 prior and the complete Stage-2 objective.

AI USE STATEMENT

Generative AI tools were used solely for language editing during the preparation of this submission. The authors reviewed all AI-assisted changes, verified the equations, citations, experimental values, and figures against the source materials, and take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

The model, two-stage training procedure, data partitions, evaluation protocol, and fixed implementation settings are described in Sections 3–4 and the appendix. The supplementary source files include the implementation and dataset details needed to reproduce the reported experiments.

REFERENCES

Fanglin Bao, Xueji Wang, Shree Hari Sureshbabu, Gautam Sreekumar, Liping Yang, Vaneet Aggarwal, Vishnu N. Boddeti, and Zubin Jacob. Heat-assisted detection and ranging. *Nature*, 619(7971):743–748, 2023. doi: 10.1038/s41586-023-06174-6. URL https://doi.org/10.1038/s41586-023-06174-6.

Zonghao Han, Ziye Zhang, Shun Zhang, Ge Zhang, and Shaohui Mei. Aerial visible-to-infrared image translation: Dataset, evaluation, and baseline. *Journal of Remote Sensing*, 3:0096, 2023. doi: 10.34133/remotesensing.0096. URL https://doi.org/10.34133/remotesensing.0096.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

Bin Hu, Chenqiang Gao, Shurui Liu, Junjie Guo, Fang Chen, Fangcen Liu, and Junwei Han. CM-Diff: A single generative network for bidirectional cross-modality translation diffusion model between infrared and visible images, 2025. URL https://arxiv.org/abs/2503.09514.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Han Huang, Yan Huang, and Liang Wang. VI-Diff: Unpaired visible-infrared translation diffusion model for single modality labeled visible-infrared person re-identification, 2023. URL https://arxiv.org/abs/2310.04122.

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. doi: 10.1109/CVPR.2015.7298706. URL https://doi.org/10.1109/CVPR.2015.7298706.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5967–5976, 2017. doi: 10.1109/CVPR.2017.632.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3489–3497, 2021. doi: 10.1109/ICCVW54120.2021.00389. URL https://doi.org/10.1109/ICCVW54120.2021.00389.

Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schrödinger bridge. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.15086.

Vladimir V. Kniaz, Vladimir A. Knyaz, Jiří Hladůvka, Walter G. Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *European Conference on Computer Vision Workshops*, pp. 606–624. Springer, 2018. doi: 10.1007/978-3-030-11024-6_46.

Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for GAN training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10651–10662, 2022. URL https://arxiv.org/abs/2112.09130.

Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. BBDM: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1952–1961, 2023. doi: 10.1109/CVPR52729.2023.00194. URL https://doi.org/10.1109/CVPR52729.2023.00194.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A. Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-image schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22042–22062. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23ai.html.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5792–5801, 2022. doi: 10.1109/CVPR52688.2022.00571. URL https://doi.org/10.1109/CVPR52688.2022.00571.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Decao Ma, Juan Su, Shaopeng Li, and Yong Xian. Aerialirgan: Unpaired aerial visible-to-infrared image translation with dual-encoder structure. *Scientific Reports*, 14(1):22105, 2024. doi: 10.1038/s41598-024-73381-0.

Fangyuan Mao, Jilin Mei, Shun Lu, Fuyang Liu, Liang Chen, Fangzhou Zhao, and Yu Hu. PID: Physics-informed diffusion model for infrared image generation. *Pattern Recognition*, 169: 111816, 2026. doi: 10.1016/j.patcog.2025.111816. URL https://doi.org/10.1016/j.patcog.2025.111816.

Catherine Ordun, Alexandra Cha, Edward Raff, Sanjay Purushotham, Karen Kwok, Mason Rule, and James Gulley. A generative approach for image registration of visible-thermal (VT) cancer faces. In *Artificial Intelligence over Infrared Images for Medical Applications*, volume 14298 of *Lecture Notes in Computer Science*, pp. 91–100. Springer, Cham, 2023. doi: 10.1007/978-3-031-44511-8_7. URL https://doi.org/10.1007/978-3-031-44511-8_7.

Mehmet Akif Özkanoğlu and Sedat Ozer. Infragan: A gan architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155:69–76, 2022. doi: 10.1016/j.patrec.2022.01.026.

Jay Nitin Paranjape, Celso M. de Melo, and Vishal M. Patel. F-ViTA: Foundation model guided visible to infrared translation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5633–5642, 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Paranjape_F-ViTA_Foundation_Model_Guided_Visible_to_Infrared_Translation_WACV_2026_paper.html.

Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pp. 319–345. Springer, 2020. doi: 10.1007/978-3-030-58545-7_19.

Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models, 2024. URL https://arxiv.org/abs/2403.12036.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Lingyan Ran, Lidong Wang, Guangcong Wang, Peng Wang, and Yanning Zhang. Diffv2ir: Visible-to-infrared diffusion model via vision-language understanding, 2025. URL https://arxiv.org/abs/2503.19012.

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.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. doi: 10.1109/TCSVT.2022.3168279. URL https://doi.org/10.1109/TCSVT.2022.3168279.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83–84:79–92, 2022. doi: 10.1016/j.inffus.2022.03.007. URL https://doi.org/10.1016/j.inffus.2022.03.007.

Teledyne FLIR. Free teledyne flir thermal dataset for algorithm training. https://oem.flir.com/solutions/automotive/adas-dataset-form/, 2018. Accessed: 2026-07-20.

Rubén Usamentiaga, Pablo Venegas, Jon Guerediaga, Laura Vega, Julio Molleda, and Francisco Bulnes. Infrared thermography for temperature measurement and non-destructive testing. *Sensors*, 14(7):12305–12348, 2014. doi: 10.3390/s140712305. URL https://doi.org/10.3390/s140712305.

Michael Vollmer and Klaus-Peter Möllmann. *Infrared Thermal Imaging: Fundamentals, Research and Applications*. Wiley-VCH, 2nd edition, 2017. doi: 10.1002/9783527693306. URL https://doi.org/10.1002/9783527693306.

Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3508–3515, 2022. doi: 10.24963/ijcai.2022/487. URL https://doi.org/10.24963/ijcai.2022/487.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Sidi Wu, Yizi Chen, Samuel Mermet, Lorenz Hurni, Konrad Schindler, Nicolas Gonthier, and Loïc Landrieu. Stegogan: Leveraging steganography for non-bijective image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7922–7931, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_StegoGAN_Leveraging_Steganography_for_Non-Bijective_Image-to-Image_Translation_CVPR_2024.html.

Jiuhong Xiao, Roshan Nayak, Ning Zhang, Daniel Tortei, and Giuseppe Loianno. Thermalgen: Style-disentangled flow-based generative models for RGB-to-thermal image translation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 33111–33133. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2f9e10c28c4052570ef9e39cc4dd69b9-Abstract-Conference.html.

Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8886–8895, 2022. doi: 10.1109/CVPR52688.2022.00868. URL https://doi.org/10.1109/CVPR52688.2022.00868.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2242–2251, 2017. doi: 10.1109/ICCV.2017.244.

Table 3: Key fixed implementation settings for the reported T2IR configuration. $\lambda^{\mathrm{P}}_{\mathrm{cyc}}$ and $\lambda^{\mathrm{P}}_{\mathrm{idt}}$ weight LPIPS in Stage 1.

Setting	Value
Backbone	One-step latent diffusion with `DDPMScheduler` (Ho et al., 2020); $T = 1000$, one scheduler timestep, update at $t_\star = 999$
Trainable adapters	UNet/VAE LoRA ranks $128/4$
Stage 1	$(\lambda_{\mathrm{cyc}}, \lambda^{\mathrm{P}}_{\mathrm{cyc}}, \lambda_{\mathrm{idt}}, \lambda^{\mathrm{P}}_{\mathrm{idt}}, \lambda_{\mathrm{gan}}) = (1.0, 1.0, 1.0, 1.0, 0.5)$; bf16
Stage 1 adversarial training	Two vision-aided discriminators (Kumari et al., 2022) with frozen CLIP ViT-B/32 backbones; three generator updates followed by fake and real discriminator updates
Stage 2	fp32
Shift and leakage settings	Sobel quantile $\eta = 0.85$
Thermal-source settings	Compact-source components use fixed area and shape criteria
Thermal-field operator	Normalized Gaussian kernel 31×31 with $\sigma = 8$

A Implementation Details

Table 3 summarizes the key fixed settings needed to reproduce the reported model. The entries of $\boldsymbol{\lambda}_r$, $\boldsymbol{\lambda}_{st}$, and $\boldsymbol{\lambda}_t$ follow the loss vectors defined in the main paper. Unless stated in an ablation, all settings are held fixed.

Ablation protocol. Setting an entry of $\boldsymbol{\lambda}_r$ or $\boldsymbol{\lambda}_{st}$ to zero removes its reconstruction or alignment-and-leakage term; setting $\boldsymbol{\lambda}_t = \mathbf{0}$ removes the source-and-field objective. The source-response and thermal-field variants set $[\boldsymbol{\lambda}_t]_{1:2} = \mathbf{0}$ and $[\boldsymbol{\lambda}_t]_{3:4} = \mathbf{0}$, respectively. Other coefficients and modules remain fixed without retuning.

A.1 Stage-1 Initialization Objective

The adopted CycleGAN-Turbo initialization objective (Parmar et al., 2024) uses unpaired samples x and y. Define $\hat{y} = F_{\mathrm{RGB}\to\mathrm{IR}}(x)$ and $\hat{x} = F_{\mathrm{IR}\to\mathrm{RGB}}(y)$. The cycle and identity terms are

$$\begin{aligned}
\mathcal{L}_{\mathrm{cyc}} &= \lambda_{\mathrm{cyc}}\big[\ell_1(F_{\mathrm{IR}\to\mathrm{RGB}}(\hat{y}), x) + \ell_1(F_{\mathrm{RGB}\to\mathrm{IR}}(\hat{x}), y)\big] \\
&\quad + \lambda^{\mathrm{P}}_{\mathrm{cyc}}\big[\ell_{\mathrm{P}}(F_{\mathrm{IR}\to\mathrm{RGB}}(\hat{y}), x) \\
&\qquad + \ell_{\mathrm{P}}(F_{\mathrm{RGB}\to\mathrm{IR}}(\hat{x}), y)\big], \\
\mathcal{L}_{\mathrm{idt}} &= \lambda_{\mathrm{idt}}\big[\ell_1(F_{\mathrm{RGB}\to\mathrm{IR}}(y), y) + \ell_1(F_{\mathrm{IR}\to\mathrm{RGB}}(x), x)\big] \\
&\quad + \lambda^{\mathrm{P}}_{\mathrm{idt}}\big[\ell_{\mathrm{P}}(F_{\mathrm{RGB}\to\mathrm{IR}}(y), y) \\
&\qquad + \ell_{\mathrm{P}}(F_{\mathrm{IR}\to\mathrm{RGB}}(x), x)\big].
\end{aligned} \tag{16}$$

Here, ℓ_{P} denotes LPIPS. The adversarial update schedule and the remaining fixed optimization settings are summarized in Table 3.

B Dataset Details

Additional acquisition details. The RGB sensor has a $3.75\ \mu\mathrm{m}$ pixel pitch and a maximum frame rate of $30\ \mathrm{frame/s}$; the infrared detector has a $12\ \mu\mathrm{m}$ pixel pitch and operates at 25–$50\ \mathrm{frame/s}$. For calibration, we capture a black-and-white board with contrast features observable in both modalities from multiple views and extract the corresponding feature points used for the spatial transformation described in the main paper.

Training pools and test-set isolation. The Stage-1 visible pool contains 92,313 images from the training splits in Table 4; its infrared pool contains the corresponding infrared training images and a filtered IR-500K (Ran et al., 2025) subset, totaling 561,182 images. IR-500K covers natural landscapes, urban environments, autonomous driving, aerial imagery, and surveillance.

Table 4: Manual training and test partitions used in our experiments.

Dataset	Train / Test samples
KAIST	16,001 / 1,778
LLVIP	13,939 / 1,549
M3FD	4,000 / 200
MSRS	1,304 / 140
VTUAV	10,304 / 1,144
AVIID	1,152 / 128
DroneVehicle	12,098 / 1,344
FLIR	4,628 / 514
THIRAlign	28,887 / 3,327

Table 5: Auxiliary-loss and thermal-group ablations evaluated on the fixed held-out THIRAlign test split.

Configuration	PSNR ↑	SSIM ↑	LPIPS ↓
Full objective	**24.18**	**0.692**	**0.243**
w/o perceptual loss	23.42	0.641	0.285
w/o shift-tolerant loss	23.31	0.666	0.273
w/o source/field statistics	23.40	0.662	0.265
w/o source-response terms	23.10	0.652	0.281
w/o thermal-field terms	23.52	0.639	0.271
w/o leakage term	23.56	0.622	0.268

Before constructing the training pools, we cross-reference IR-500K against the predefined evaluation partitions and remove infrared images associated with KAIST, LLVIP, and the other evaluation test sets.

C Auxiliary Loss Ablations

Table 5 reports controlled ablations on the fixed held-out THIRAlign test split. All variants use the same Stage-1 checkpoint, paired data, training schedule, trainable modules, preprocessing, optimizer, and evaluation protocol.