TherA: Thermal-Aware Visual-Language Prompting for Controllable RGB-to-Thermal Infrared Translation

Supplementary Material

A. Additional Implementation Details

A.1. Datasets

R2T2. R2T2 is a RGB–TIR–text dataset that we designed specifically to train TherA-VLM and our diffusion-based translation model. R2T2 provides spatially aligned RGB–TIR image pairs together with canonicalized thermal-aware captions that are used to train TherA-VLM and the VLM-conditioned diffusion model (see Sec. B for curation details). In total, R2T2 contains 112,970 RGB–TIR–text triplets. We use all available RGB–TIR pairs to train our models, with the exception of samples that overlap with the VTMOT-test, FLIR, and M3FD sets. R2T2 serves as the primary supervision source for both TherA-VLM and our diffusion-based translation model, and all models used in our zero-shot experiments are trained exclusively on R2T2. Example data samples are provided in Fig. 7

Table 4. Constituent datasets of R2T2 and their statistics. "Sample Number (Translation)" denotes the number of RGB–TIR pairs used for training the translation model. The total across all sources is 112,970 pairs.

Dataset	Sample Number (Translation)	Average Resolution	Viewpoint	Split
AVIID [34]	683	480×480	Aerial	Train
CAMEL [36]	3481	Diverse	CCTV	Train
FLIR [18]	5141	640×512	Driving	Finetune/Test
KAIST [31]	6457	640×512	Driving	Train
LLVIP [32]	15488	640×512	CCTV	Train
M3FD [3]	4200	512×384	Driving	Finetune/Test
METU-VisTIR [30]	408	640×512	Ego-View	Train
MSRS [35]	1340	640×480	Driving	Train
NSAVP [38]	23520	640×512	Driving	Train
SMOD [33]	4553	640×512	Driving	Train
VIVID [37]	23012	640×512	Driving	Train
VTMOT [6]	25079	Diverse	Ego-view	Train/Val

Table 4 summarizes the constituent datasets used to build R2T2, including the number of RGB–TIR pairs contributing to translation training, their typical resolutions, viewpoints (driving, CCTV, aerial, ego-view), and data split. This composition provides a diverse mixture of viewpoints and environments, while ensuring that samples from FLIR, M3FD, and the VTMOT-test split used for benchmark finetuning and evaluation are clearly separated from those used for training.

Benchmark datasets and evaluation protocol. For benchmark finetuning and evaluation, we use the M3FD [3], FLIR [18], and CART [51] datasets. Following the non-overlapping splits provided in prior work [15, 59], we use 3,550 training images and 650 test images for M3FD, and 4,128 training images and 1,013 test images for FLIR. For CART, we adopt the official test split of 280 images from [51] and do not use any CART images for training.

For M3FD and FLIR, we consider two evaluation regimes: (i) *zero-shot*, where the model is trained only on R2T2 and evaluated directly on the corresponding test split, and (ii) *finetuned*, where the R2T2-pretrained model is further finetuned on the respective training split and evaluated on the held-out test split. For CART, we report only zero-shot results: the model is trained on R2T2 and evaluated on the CART test split without any additional finetuning.

A.2. TherA-VLM

We fine-tune LLaVA v1.5 (7B) [40] with a CLIP ViT-L/14-336 vision encoder [60] for thermally aware instruction tuning. The model is trained on the R2T2 dataset with a 70:30 train/validation split. We use the Vicuna v1 conversation template, and compute a standard next-token cross-entropy loss over assistant tokens while masking out image tokens and user tokens from the loss.

For fine-tuning, we attach LoRA adapters (rank 128, $\alpha = 256$) to the linear layers of the language model. The LLM backbone and vision encoder remain frozen; only the LoRA parameters and the image-to-text projector weights are updated. We optimize with AdamW (weight decay 0.0), using a cosine learning-rate schedule with a warmup ratio of 0.03 and a maximum sequence length of 2048 tokens. The base learning rate for the language model is set to 2×10^{-4} and the projector learning rate to 2×10^{-5}. Training is run for one epoch with a per-GPU batch size of 16 and gradient accumulation of 2 across 4 GPUs, resulting in an effective batch size of 128.

A.3. Image Translation

For image translation, input RGB and TIR images are resized to 256×256, converted to tensors, and normalized to $[-1, 1]$. We use the Stable Diffusion v1.4 VAE (frozen) to encode both RGB and TIR frames into latent representations. The UNet operates on an 8-channel input formed by concatenating noisy TIR latents with RGB latents along the channel dimension. We initialize the UNet weights from Stable Diffusion v1.4 and the text-to-image adapter (TE-Adaptor) from the LLaMA–UNet adapter of [42]. The VAE and the VLM model remain frozen; we update the UNet and TE-Adaptor parameters.

We train the diffusion model with AdamW on the UNet

and adapter parameters, using decoupled weight decay (10^{-4} on weight tensors), betas $(0.9, 0.999)$, and $\epsilon = 10^{-8}$. We apply a cosine learning-rate schedule with warm-up and set the peak learning rate to 1×10^{-4}. Training is performed for 100 epochs with a per-GPU batch size of 50 and gradient accumulation of 1 on 4 NVIDIA A6000 GPUs. We use `bfloat16` mixed precision with `TF32` enabled and apply gradient clipping with a maximum global norm of 0.5. The training objective is the standard diffusion noise-prediction MSE loss. During training, we employ classifier-free guidance (CFG) dropout by independently dropping the unconditional, text, and image conditions with probability 0.1 to improve robustness.

B. R2T2 Curation

B.1. Thermal-aware Text Generation

B.1.1. From Radiometric Chain to Semantic Schema

Although we only observe thermal images that display relative temperature in our datasets (8-bit, normalized TIR frames that preserve the ordering of temperatures, but not their absolute values), the formation of these images is still governed by the standard radiometric chain [14]. The radiometric chain states that the irradiance measured by an TIR detector can be written as

$$\Phi_{\text{det}} = \tau_{\text{atm}}\, \varepsilon\, \Phi^{bb}(T_{\text{obj}}) + \tau_{\text{atm}}(1-\varepsilon)\, \Phi_{\text{amb}} + (1-\tau_{\text{atm}})\, \Phi_{\text{atm}}, \tag{8}$$

where ε is the material emissivity, T_{obj} is the object temperature, $\Phi^{bb}(\cdot)$ denotes black-body emission, and $T_{\text{amb}}, T_{\text{atm}}, \tau_{\text{atm}}$ control ambient and atmospheric contributions.

Because our TIR images are normalized, we cannot directly recover absolute values of ε, T_{obj}, or T_{amb} from pixel intensities. However, the *relative* contrast and structure observed in TIR still arise from variations in these physical factors. Our canonical schema is designed as a high-level, discretized parameterization of this radiometric chain:

- **Material** groups objects into emissivity classes, providing a coarse proxy for ε.
- **Heat Emission States** encodes the relative thermal state of an object (e.g., parked vs. driving car), acting as a proxy for T_{obj}.
- **Scene** (time of day, weather, environment) captures conditions that influence $T_{\text{amb}}, T_{\text{atm}}$ and τ_{atm} (e.g., solar loading on roads, nighttime cooling).
- **Object** category supplies priors over typical spatial heat distributions (e.g., hot engine region vs. cooler body, warm human torso vs. cooler background).

Without these four components, key terms in the radiometric chain become effectively unobserved from RGB alone, making the RGB-to-TIR mapping severely underdetermined. By explicitly modeling these components, we obtain a physically motivated, semantically interpretable discretization of the radiometric chain that provides exactly the information needed for stable conditioning of the diffusion model, while remaining compatible with the relative (normalized) TIR images used in practice.

B.1.2. Canonical Schema Generation

Generating text only from an RGB image does not reveal how the scene appears in the thermal infrared domain. Since our goal is to train a *thermal-aware* vision–language model, we construct supervision using a multimodal reasoning model, Gemini 2.5 Pro [29], tied explicitly to our radiometric schema (Sec. B). Given paired RGB–TIR inputs, we query Gemini with a structured, slot-based prompt that returns three fields: `SCENE`, `OBJECTS`, and `PRIOR`. The prompt constrains object names, materials, colors, and states to a fixed vocabulary, enforces a specific format (e.g., `name(material=...; color=...; position=...; state=active|passive)`), and asks Gemini to report only high-confidence thermal attributes. The `SCENE` and `OBJECTS` fields describe time of day, weather, environment, object categories, materials, and heat emission (active/passive) states, whereas `PRIOR` is reserved for thermal context that is typically not recoverable from RGB alone (e.g., solar loading on pavement, whether parked cars are active or passive). The full prompt is available in Fig. 9

Handling non-observable thermal states: PRIOR. Even with paired RGB–TIR inputs, some thermal factors are ambiguous from RGB appearance (e.g., accumulated solar loading, or whether a vehicle is currently running). If these factors are omitted from the conditioning text while they are clearly present in the TIR image, the model is trained under a semantic–image mismatch, which encourages either hallucinated thermal patterns or ignoring the text altogether. To mitigate this, we introduce the `PRIOR` slot, which is instantiated from TIR evidence during data curation (e.g., "sun-heated pavement", "parked cars passive", "electronics off"). During training and inference, we append the PRIOR phrases to the user instruction that conditions TherA-VLM, leveraging LLaVA's conversational interface. In practice, PRIOR serves as a physically motivated control knob: it aligns the conditioning text with the ground-truth TIR during training and later enables explicit text-guided control of thermal behaviour in the translation model.

Canonicalizing free-form reasoning. Raw free-form descriptions from Gemini contain a large, redundant, and noisy vocabulary, which led to unstable conditioning in early experiments. We therefore canonicalize Gemini's reasoning into a compact, physics-informed schema. Scene

[SCENE]: outdoor, urban, night
[OBJECTS]: **person**(material=skin, fabric; color=white; position=left; state=active)
car{1}(material=metal, plastic; color=white; position=top; state=active)
car{2}(material=metal, plastic; color=white; position=center; state=passive)
car{3}(material=metal, plastic; color=black; position=left; state=passive)
car{4}(material=metal, plastic; color=black; position=bottom; state=passive)
bicycle(material=metal, rubber; color=yellow; position=top; state=passive)
pavement(material=masonry; color=gray; position=bottom; state=passive)
vegetation(material=organic; color=green; position=left; state=passive)
[PRIOR]: active car headlights/engine are prominent heat sources

(a) RGB-TIR-text example from LLVIP [32]

[SCENE]: outdoor, park, daytime, solar-loading
[OBJECTS]: **person{1}**(material=skin, fabric; color=white; position=left; state=active)
bicycle{1}(material=metal, plastic; color=gray; position=left; state=passive)
pavement(material=masonry, paint; color=gray; position=bottom; state=passive)
barrier(material=organic; color=green; position=center; state=passive)
vegetation(material=organic; color=green; position=center; state=passive)
[PRIOR]: strong solar loading on pavement

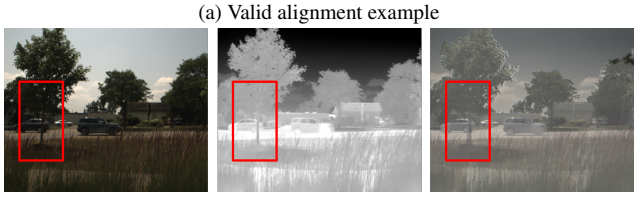
(b) RGB-TIR-text example from SMOD [33]

[SCENE]: outdoors, daytime, sunny, roadway, aerial-view
[OBJECTS]: **car{1}**(material=metal, glass; position=left; state=active),
Person{1}(material=skin, fabric; position=right; state=active),
pavement(material=other; position=center; state=passive),
vegetation(material=organic; position=left, right; state=passive)
[PRIOR]: solar-loading on pavement, active vehicle engine

(b) RGB-TIR-text example from AVIID [34]

Figure 7. Examples of RGB–TIR pairs in R2T2 dataset.

(a) Valid alignment example



(b) Invalid alignment example

Figure 8. **Examples of pseudo-aligned RGB–TIR pairs.** We use alpha-blended overlays (right) of RGB (left) and TIR (middle) to visually screen for residual misalignment, and keep only pairs with minimal visible parallax.

and object labels are mapped to the COCO *things/stuff* taxonomy [61], and material groups are mapped to emissivity classes defined in HADAR [62]. Each object instance is annotated with a canonical material group, color, position (left/right/top/bottom), and heat emission state (*active* or *passive*), yielding a vocabulary of 23 object classes, 13 material groups, 14 colors, 4 positions, and 2 heat emission states. We first instruct Gemini to follow this schema and then apply a post-hoc LLM pass that merges synonyms into canonical tokens (e.g., "sedan"/"SUV" $\rightarrow$ `car`; "asphalt"/"road surface" $\rightarrow$ `road`). Finally, because we operate on RGB–TIR pairs, we run a consistency check in which the LLM flags and corrects contradictions between the canonicalized text and the observed TIR patterns (e.g., an object described as "cool" but appearing hotter than its surroundings). Although PRIOR and active/passive states may still be occasionally mispredicted, the restricted vocabulary and consistency filtering substantially reduce noise and yield stable, thermally grounded captions, which we find to be markedly more effective than raw free-form descriptions in our ablations.

```
TASK: Describe how the RGB scene would appear in TIR. Use only high-confidence facts; abstain
otherwise.
INPUTS: RGB_IMAGE, THERMAL_IMAGE
OUTPUT: Single text block.  No extra text outside fields.
SCENE: Identify the general environment of the weather that could influence the thermal
characteristics of the image.  These could include locations, weather, time of the day,
indoor/outdoors, season.  Refer to heat intensities of the thermal image and the appearance of
RGB images to make reasonable deductions.  Only output deductions that you are confident with.
State them by keywords separated by commas.  Omit uncertain items.
OBJECTS: ITEM FORMAT: name(material=...; color=...; position=left|right|top|bottom|center;
state=active|passive).
THINGS (instanceable):  {person, people, car, truck, motorcycle, bicycle, other_vehicle,
building_equipment, streetlight, sensor, furniture, electronics, object, animal, potted_plant}
STUFF (non-instanceable; record ONCE only):  {building, pavement, barrier, structure,
street-furniture, vegetation, traffic-sign}
MATERIALS (allowed):  {metal, glass, plastic, rubber, masonry, organic, fabric, skin, air, paint,
light, water, other}
COLORS (allowed):  {white, black, gray, brown, cream, red, blue, yellow, green, orange, purple,
pink, navy, na}
STATE RULES: active = self-heating (engines/electronics/living bodies); passive =
environment/solar only.  Vehicles:  active if hotspot in hood/exhaust/wheels; side-parked + no
hotspot -> passive.
GROUPING (THINGS only):  split_by_state -> if (center_dist < 0.05*diag OR IoU > 0.10 OR occlusion)
-> plural (cars, people) else index as name{1}, name{2}, ...
STUFF: no indices unless that part is a Prominent heat source -- OMIT: small & low-contrast &
thermally irrelevant -- UNCERTAIN but relevant:  map to nearest allowed name; material=other if
unsure -- EXCLUDE: artifacts/flares/blurs from OBJECTS
PRIOR: - Short phrases of thermal context NOT visible from RGB (e.g., solar-loading, sky<<ground,
parked cars passive, electronics off).  - Mention active/passive classes if relevant.
Notes:  - Use ONLY allowed tokens for names/materials/colors - Prefer omission or "other" over
speculation - Output exactly in the field order above
```

Figure 9. **Gemini Prompt.** Prompt used to query Gemini 2.5 Pro for thermal-aware captions over our schema fields (SCENE, OBJECTS, PRIOR).

B.2. Pseudo-aligned RGB–TIR Pairs

Many public RGB–TIR datasets are time-synchronized but not spatially aligned, which makes them unsuitable for direct pixel-level supervision. To increase the diversity of our training data, we convert such datasets into *pseudo-aligned* RGB–TIR pairs. Our goal is not to obtain perfect pixel-wise registration, but to identify frames where the residual misalignment is small enough to be useful for training. These pseudo-aligned pairs are used only for training the translation model; all quantitative evaluations rely exclusively on natively aligned benchmarks (M3FD, FLIR, CART).

For each synchronized RGB–TIR frame, we estimate cross-modal correspondences with MINIMA [10], and fit a global warp that maps the RGB image into the TIR view. For datasets where camera extrinsics and LiDAR scans are available, we first apply the provided calibration to coarsely rectify the views and then run the same correspondence-based refinement. Because the scenes are not strictly planar and the sensors have a non-zero baseline, the resulting pairs are only approximately registered; we therefore explicitly treat them as pseudo-aligned.

To filter out pairs with unacceptable parallax or warping artifacts, we adopt a simple human-in-the-loop check. For each candidate pair, we generate an alpha-blended RGB–TIR overlay (0.5/0.5) and ask three annotators to label it as *accept* or *reject*, using a small set of reference examples for guidance (see Fig. 8). Only pairs with majority *accept* votes are retained. Applying this procedure across several RGB–TIR datasets [36–38] yields 50,013 additional pseudo-aligned pairs used solely as extra training data.

[SCENE]: outdoors, daytime, overcast, driving-view
[OBJECTS]: **car{1}**(material=metal; color=gray; position=right; state=passive),
car{2}(material=metal; color=black; position=center; state=active),
car{3}(material=metal; color=white; position=left; state=active),
car{4}(material=metal; color=black; position=center; state=active),
car{5}(material=metal; color=black; position=center; state=active),
building(material=masonry; color=cream; position=right; state=passive),
pavement(material=masonry; color=gray; position=center; state=passive),
vegetation(material=organic; color=green; position=top; state=passive),
traffic-sign(material=metal; color=blue; position=left; state=passive)

(a) Example TherA-VLM Output: Cityscapes [53]

[SCENE]: outdoors, urban, rain, overcast, ego-view
[OBJECTS]: **people**(material=skin; color=white, black; position=center; state=active),
person{1} (material=skin; color=red; position=right; state=active),
building(material=masonry; color=gray; position=top; state=passive),
pavement(material=masonry; color=gray; position=bottom; state=passive),
vegetation(material=organic; color=green; position=left; state=passive),
streetlight(material=metal; color=gray; position=left; state=passive)

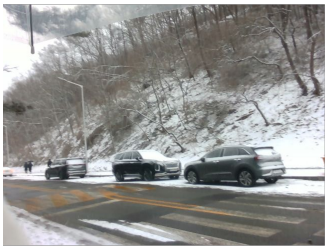
(b) Example TherA-VLM Output: FMB [63]

[SCENE]: outdoors, daytime, sunny, solar_loading, drone-view
[OBJECTS]: **people**(material=skin; color=na; position=center; state=active),
object(material=other; color=white; position=center; state=passive),
pavement(material=masonry; color=gray; position=center; state=passive),
vegetation(material=organic; color=green; position=center; state=passive),
water-body(material=water; color=blue; position=top; state=passive),

(c) Example TherA-VLM Output: CART [51]

[SCENE]: indoors, daytime, ego-view
[OBJECTS]: **person{1}**(material=skin; color=yellow; position=left; state=active),
person{2}(material=skin; color=black; position=center; state=active),
object(material=plastic; color=white; position=left; state=passive),
furniture(material=plastic; color=black; position=center; state=passive),
furniture(material=plastic; color=gray; position=bottom; state=passive),
building(material=masonry; color=white; position=top; state=passive),
structure(material=glass; color=na; position=right; state=passive)

(d) Example TherA-VLM Output: Trimodal [64]



[SCENE]: outdoors, daytime, winter, snow, driving-view
[OBJECTS]: **car{1}**(material=metal; color=black; position=center; state=passive),
car{2}(material=metal; color=gray; position=right; state=passive),
car{3}(material=metal; color=black; position=left; state=passive),
streetlight(material=metal; color=gray; position=left; state=passive),
pavement(material=masonry; color=gray; position=bottom; state=passive),
vegetation(material=organic; color=brown; position=top; state=passive)

(e) Example TherA-VLM Output: Custom

Figure 10. **TherA-VLM Examples.** Example output results for TherA-VLM across different benchmarks

C. Additional Qualitative Results

C.1. TherA-VLM Examples

Figure 10 illustrates the generated language output of TherA-VLM on unseen images. Here, images from diverse scenes, including urban driving, aerial terrain, rainy streets, snowy roads, and indoor environments are presented. Across all cases, the model produces coherent SCENE descriptions (e.g., rain, snow, solar loading, indoor vs. outdoor) and assigns materials and active/passive states to objects (vehicles, people, vegetation, buildings, furniture) in a manner consistent with typical thermal behavior, providing physically grounded conditioning signals for the translation model.

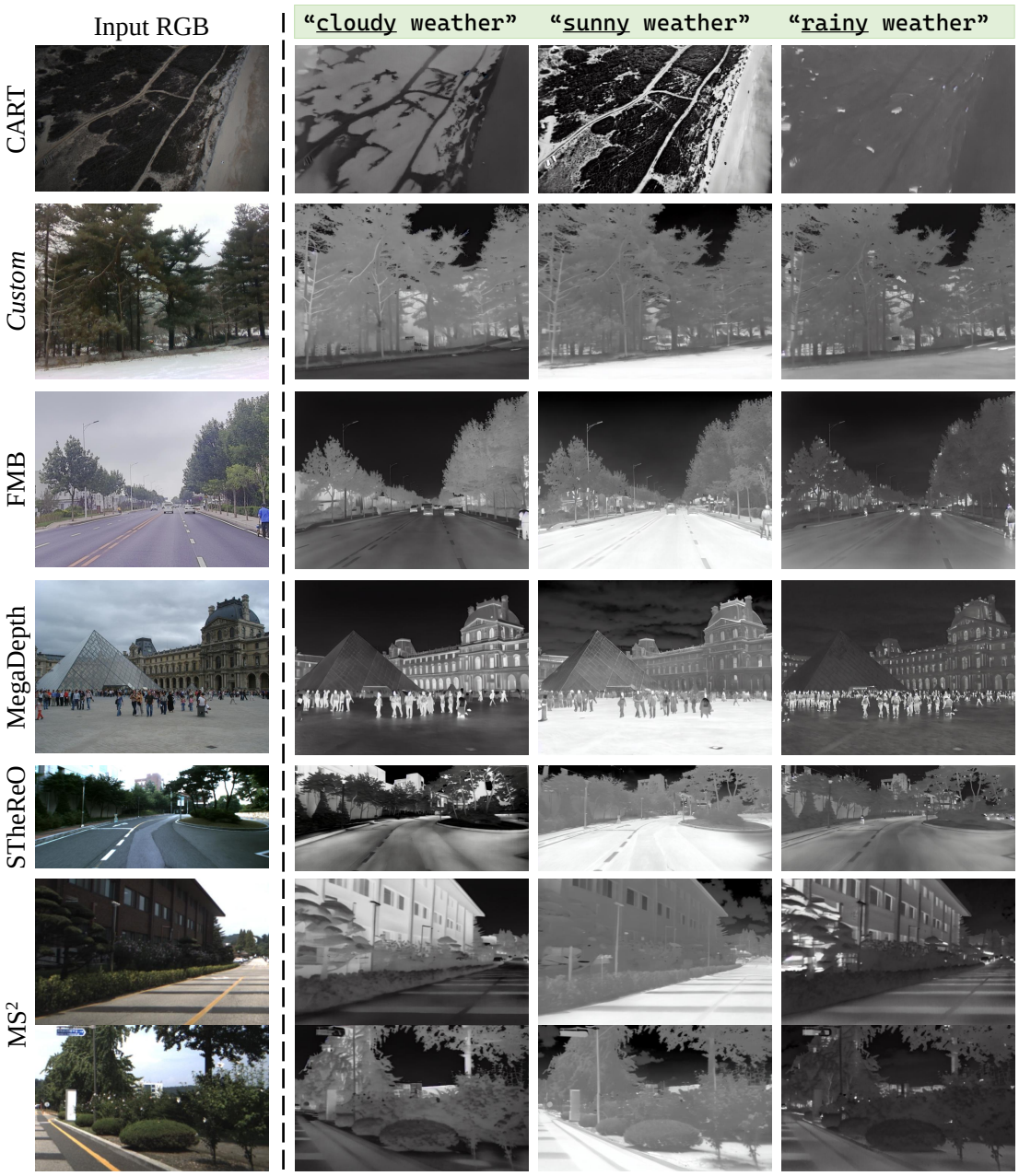


Figure 11. **Text-guided controllability** Qualitative results of *cloudy, sunny, rainy* weather on different datasets: CART [51], custom dataset, FMB [63], MegaDepth [55], MS2 [65], STheReO [57]. MS2 results were horizontally split due to its long aspect ratio.

C.2. Controllability Qualitative Results

C.2.1. Text-guided Control Qualitative Results

In Fig. 11, we present additional qualitative examples of text-guided controllability. For each RGB input, we generate translations conditioned on different weather prompts across various public benchmark datasets and a custom snowy scene, all in a zero-shot setting with diverse fields of view, seasons, and locations. We observe that the translation consistently adapts to the requested environmental condition. For example, in the custom snowy image, the apparent temperature of the ground terrain changes with the weather prompt, a behavior that is absent in existing RGB-to-TIR models. This controllability is enabled by the user-interactive TherA-VLM, which allows us to inject prior thermal conditions into the prompt that would not be recoverable from the RGB image alone.

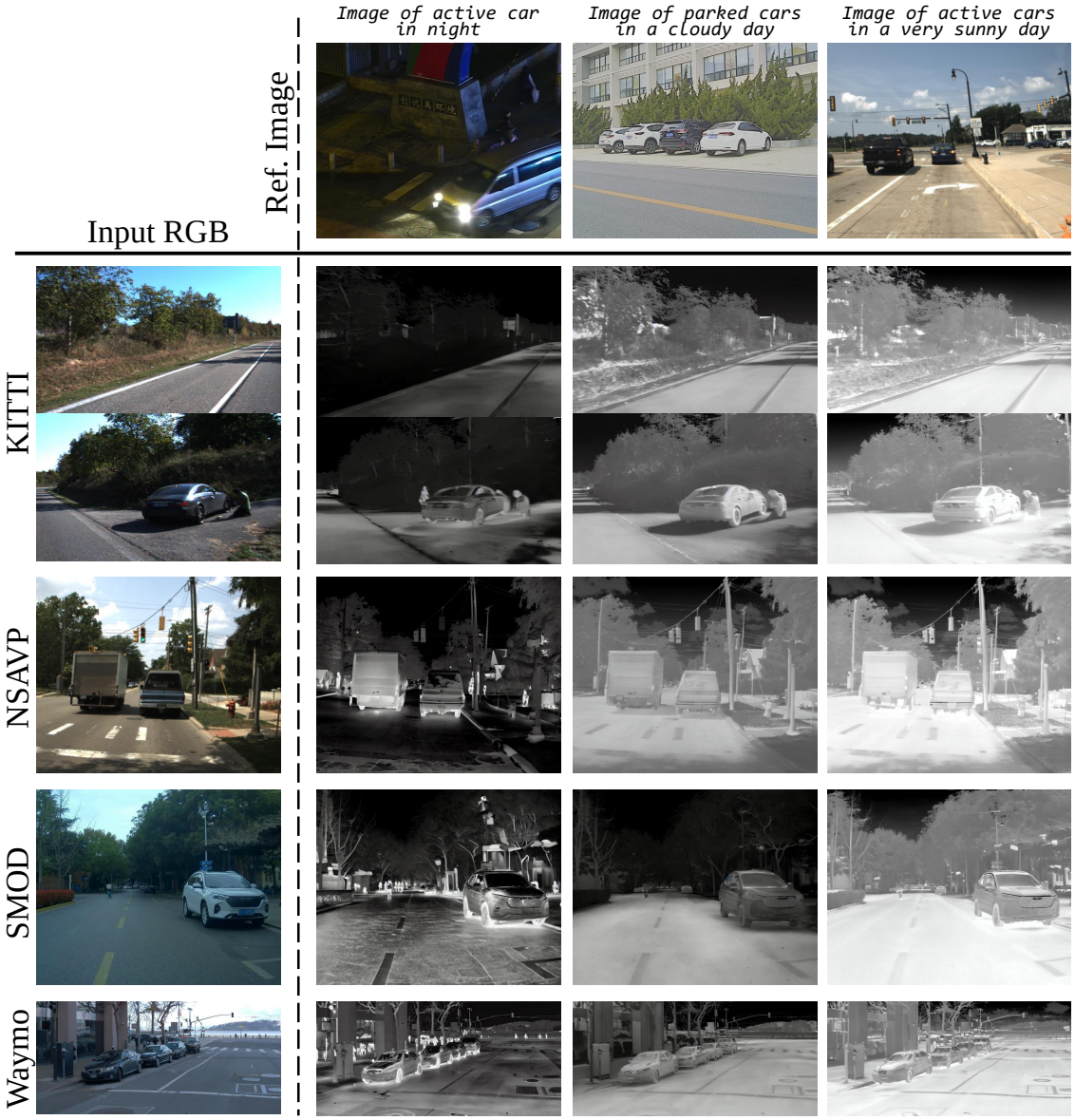


Figure 12. **Reference-guided controllability** Qualitative results of reference-guided control with reference RGB images from LLVIP [32], M3FD [3], NSAVP [38] on translating images from KITTI [66], NSAVP [38], SMOD [33], Waymo [2]. KITTI results were horizontally split due to its long aspect ratio.

C.2.2. Reference-guided Control Qualitative Results

We also provide additional examples of reference-guided controllability in Fig. 12. For each input RGB image, we condition the translator on a reference RGB depicting an active car at night, parked cars on a cloudy day, or active cars on a sunny day. The translated TIR images adapt global appearance and object-level heat patterns to follow the thermal characteristics implied by the reference. With the night-time active-car reference, vehicles become strong hotspots against a cooler background; with the parked-car reference, vehicles appear cooler and closer to the background; with the sunny-day reference, road surfaces and buildings brighten from solar loading while active vehicles remain highlighted. These results indicate that conditioning through TherA-VLM allows the diffusion model to inherit scene-wise illumination and object-wise emission cues from the reference image, even when the input and reference come from different datasets.

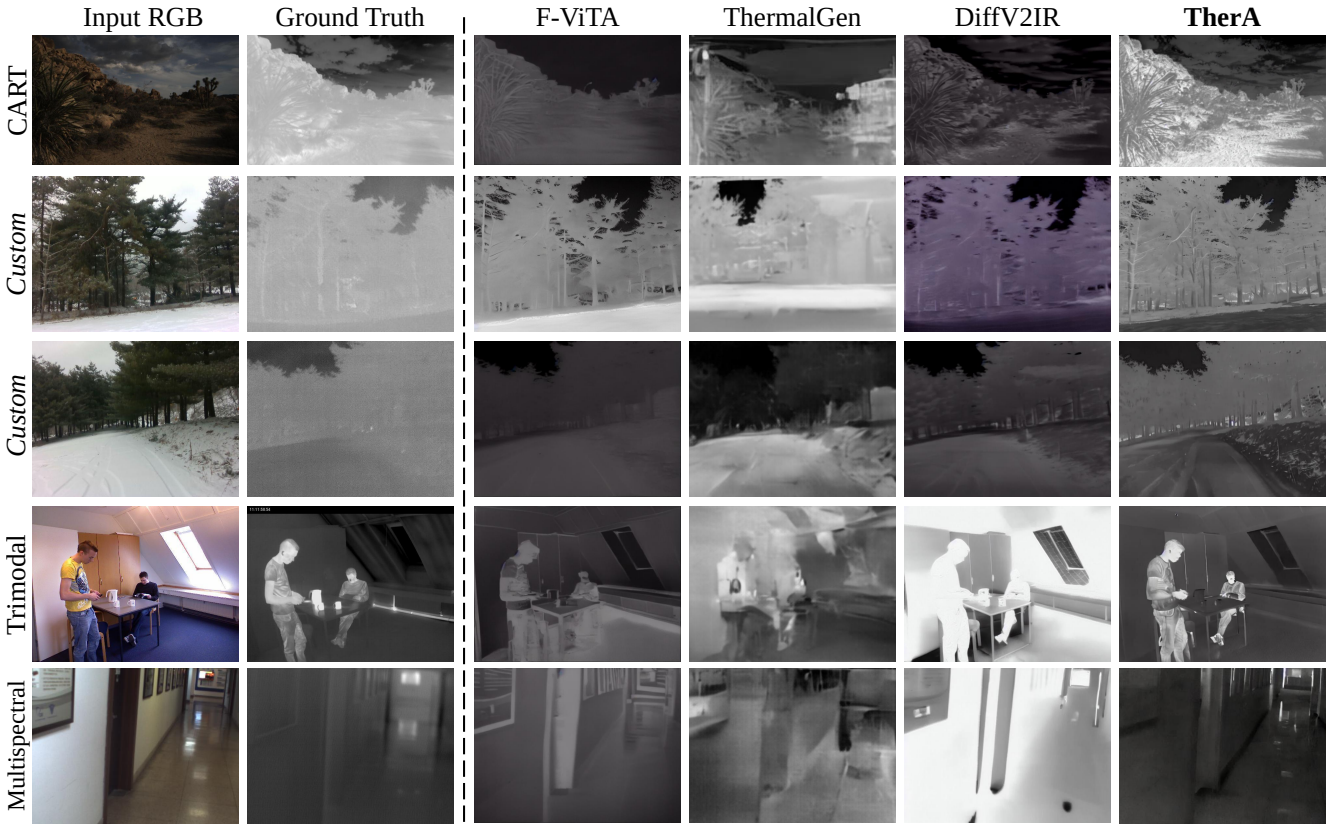


Figure 13. **Zero-shot Results** Qualitative results on RGB-TIR datasets: CART [51], custom dataset, Trimodal [64], Multispectral Motion Dataset [67].

C.3. Zero-shot Qualitative Results

We show additional translation results of TherA in Fig. 13 and Fig. 14.

In Fig. 13, we compare zero-shot translations on paired RGB–TIR datasets where ground-truth TIR is available (CART [51], custom scenes, Trimodal [64], Multispectral Motion [67]). Baseline models (F-ViTA, ThermalGen, DiffV2IR) frequently exhibit artifacts and thermally implausible outputs. For instance, in the snowy examples (second and third rows), F-ViTA produces unnaturally hot ground despite visible snow cover, and ThermalGen introduces severe ghosting and background bleeding. DiffV2IR yields more stable results but still struggles in indoor scenes, where walls and furniture appear over-heated and fine structural details are washed out. In contrast, TherA produces TIR images that more closely match the ground truth, preserving the relative temperature ordering between snow and vegetation, people and background, and corridor structures.

In Fig. 14, we evaluate true zero-shot generalization on RGB-only datasets (Cityscapes, NuScenes, MegaDepth, KITTI, Waymo). Without any paired supervision from these domains, most baselines show a noticeable degradation in quality: contrast is reduced, semantic boundaries become blurry, and expected thermal behavior is often violated (e.g., distant parked vehicles rendered as overly bright moving hotspots, or façades appearing warmer than sunlit roads). While DiffV2IR remains the strongest baseline, its translations are still softer and less physically consistent than those of TherA. TherA maintains sharper details and more realistic thermal distributions, such as cooler skies, thermally structured building façades, and appropriately warm roads and vehicles under traffic scenes.

Figure 14. **Zero-shot Results** Qualitative results on RGB-only datasets: Cityscapes [53], NuScenes [54], MegaDepth [55], KITTI [66], Waymo [2]. KITTI results were horizontally split due to its long aspect ratio.

D. Ablation Studies

D.1. CFG Scales

Table 5. Ablation results for CFG scales. Best results are highlighted in **bold** and second best are underlined.

CFG Scale		PSNR(↑)	SSIM(↑)	FID(↓)	LPIPS(↓)
Thermal VLM	Image control				
0.5	0.5	19.46	0.66	87.17	0.21
	1.5	18.67	0.64	97.60	0.23
	2.5	18.29	0.62	102.69	0.24
1.5	0.5	**19.54**	**0.67**	**87.08**	**0.21**
	1.5	18.78	0.64	96.55	0.22
	2.5	18.38	0.62	101.78	0.23
2.5	0.5	19.52	0.66	87.68	0.21
	1.5	18.86	0.64	96.36	0.22
	2.5	18.46	0.62	100.93	0.23

We analyze the impact of the dual CFG scales for image guidance (s_V) and VLM guidance (s_S) in Table 5. The results reveal a clear and significant trend. First, increasing the image guidance (s_V) consistently degrades performance across all metrics. For instance, at $s_S = 1.5$, increasing s_V from 0.5 to 2.5 drops PSNR from 19.54 to 18.38 and worsens FID from 87.08 to 101.78. Conversely, the VLM guidance (s_S) demonstrates an optimal ratio with all metrics peaking at $s_S = 1.5$.

This analysis shows that optimal translation is achieved by minimizing reliance on the raw RGB latent ($s_V = 0.5$) and assigning more weights on the thermally aware VLM guidance ($s_S = 1.5$). This confirms TherA-VLM's thermal embedding provides a more robust and physically grounded signal than the RGB latent alone. We use $s_V = 0.5$ and $s_S = 1.5$ for all experiments.

D.2. Downstream Evaluation

D.2.1. Thermal Image Segmentation

To assess the practical utility of our translations, we evaluate their impact on downstream thermal semantic segmentation. We generate pseudo-TIR data by translating Cityscapes [53] with each RGB-to-TIR model and train SegFormer [68] on the resulting images. We then evaluate this model in two controlled settings: (1) *Zero-shot* transfer to FMB [63] and MFNet [69], and (2) *Fine-tuned*, where the same SegFormer is further trained on the real FMB/MFNet training sets. Apart from the translation model used to synthesize the Cityscapes pseudo-TIR data, all factors (architecture, training schedule, and labeled target data) are kept fixed. We also include a "Real TIR" baseline trained only on the real FMB/MFNet training images, without any synthetic pre-training. For training, we kept all hyperparameters identical to the original implementation, except that we used a batch size of 4 and trained for 100 epochs.

As summarized in Tab. 6, TherA consistently yields the best segmentation performance. In the zero-shot setting, TherA achieves the highest mIoU on both datasets, reaching 27.14 on FMB and 30.43 on MFNet; on MFNet, this corresponds to a +5.05 absolute mIoU gain over the next-best model (ThermalGen). After fine-tuning on real data, models pre-trained on TherA's pseudo-TIR data still retain a clear advantage: our approach obtains 52.05 mIoU on FMB and 68.60 mIoU on MFNet, outperforming all other translation baselines and even the Real TIR baselines trained from scratch (42.04 and 57.09 mIoU, respectively).

Table 6. Quantitative comparison of class-free TIR semantic segmentation **mIoU results** [%] on FMB [63] and MFNet [69] datasets. *Zero-shot*: SegFormer trained on the Cityscapes dataset [53] translated using each model. *Fine-tuned*: The models further fine-tuned on the FMB or MFNet datasets. Best results are highlighted in **bold** and second best are underlined.

Method	**FMB** [63]		**MFNet** [69]	
	Zero-shot	Fine-tuned	Zero-shot	Fine-tuned
Real TIR	*N/A*	42.04	*N/A*	57.09
F-ViTA [16]	26.25	50.81	21.98	64.80
ThermalMGen [17]	23.49	50.19	25.38	66.24
DiffV2IR [15]	26.17	50.88	23.75	64.38
TherA	**27.14**	**52.05**	**30.43**	**68.60**

These results suggest that, under a controlled segmentation pipeline, TherA produces pseudo-TIR data that is both more informative and more physically plausible for downstream learning than alternative RGB-to-TIR translation models, leading to stronger zero-shot performance and more effective fine-tuning on limited real-world thermal data. In particular, the fact that TherA pre-training even surpasses training directly on real TIR indicates that our pseudo-TIR images encode thermal contrast patterns that are easier to learn from while remaining consistent with the statistics of real thermal imagery.

D.2.2. RGB–TIR Image Matching

We further study how different translation models affect downstream RGB–TIR perception by experimenting with cross-modal image matching. Starting from the RGB-only MegaDepth [55] dataset, we generate pseudo-TIR images using each RGB-to-TIR translator and train the same LoFTR matcher [70] on the resulting RGB–pseudo-TIR pairs. As a reference, we also include a *Baseline (RGB)* model where LoFTR is trained only on original RGB–RGB pairs from MegaDepth (no translation) and then directly applied to RGB–TIR matching. In all cases, only synthetic data is used for training; no real thermal images or real RGB–TIR pairs from the target dataset are seen during training. We then evaluate the learned matchers *zero-shot* on real RGB–TIR pairs from the METU-VisTIR dataset [30], reporting pose-estimation AUC at $5°$, $10°$, and $20°$. Similar to our segmentation experiments, all components of

the matching pipeline (network architecture, loss, training schedule, and evaluation protocol) are kept fixed; the only varying factor is the translation model used to synthesize the training TIR images. For training, we kept all hyperparameters identical to the original implementation, except that we used a batch size of 8 and trained for 25 epochs.

Table 7 summarizes the results. TherA yields the best performance across all thresholds, achieving AUC@5° = 14.98, AUC@10° = 28.98, and AUC@20° = 45.24, compared to the next-best DiffV2IR with 11.83, 26.03, and 43.17, respectively. In contrast, F-ViTA and ThermalGen provide only marginal gains over the RGB baseline: their AUC scores differ from *Baseline (RGB)* by at most ∼2–3 points, and even degrade performance at some thresholds (e.g., F-ViTA at 5°). This indicates that their pseudo-TIR outputs do little to close the RGB–TIR modality gap for matching. By comparison, DiffV2IR and especially TherA deliver large improvements over the RGB baseline, showing that physically plausible translation is crucial for learning robust cross-modal correspondences. Because all components of the matching pipeline are held fixed and only the synthetic TIR translator is varied, these gains provide further evidence that TherA's pseudo-TIR images encode more physically meaningful cross-modal cues rather than merely improving perceptual image quality.

Table 7. Quantitative comparison of RGB-TIR image matching AUC results [%] on METU_VisTIR [30] dataset. Best results are highlighted in **bold** and second best are underlined.

Method	LoFTR		
	AUC@5°	AUC@10°	AUC@20°
Baseline (RGB)	5.44	12.58	24.28
F-ViTA [16]	4.93	12.73	25.32
ThermalGen [17]	5.81	14.24	27.65
DiffV2IR [15]	11.83	26.03	43.17
TherA	**14.98**	**28.98**	**45.24**

D.3. Comparison with Reference-guided Models

We evaluated the reference-guided image translation performance of TherA against state-of-the-art example-guided image-to-image translation models, including CSGO [46], StyleID [47], and StyleSSP [48]. For all three baselines, we use the official implementations and publicly released pre-trained weights. For CSGO, we adopt the reference-guided setting by feeding the RGB input as the content image and the reference TIR frame as the style image. For StyleID and StyleSSP, we evaluate them in their recommended training-free configuration, using the RGB frame as the source and the TIR frame as the style exemplar. All methods receive exactly the same RGB–TIR pairs and reference images; only slight exception exists for TherA which leverages the RGB reference image as the reference condition in contrast to other methods which leverage TIR images as the reference image for translation. But in the end, only the translation module is varied while the evaluation protocol is kept fixed.

As shown in Fig. 15, baseline methods struggle to effectively bridge the modality gap between the visible and thermal domains. Specifically, CSGO exhibits domain-inconsistent artifacts and residual chromatic noise, failing to adequately suppress RGB color information. StyleID and StyleSSP suffer from over-smoothing and fail to capture distinct thermal physics. As a result, these baselines often fail to consistently map semantic classes to appropriate thermal intensities, frequently rendering heat-emitting objects such as pedestrians or vehicle engines with low contrast.

In contrast, TherA produces the most perceptually realistic thermal imagery: it successfully disentangles RGB texture from structure, accurately synthesizes thermal infrared appearance for distinct semantic objects, and faithfully adheres to the textural style and polarity of the reference guidance. For example, in the first row, all other methods incorrectly depict the white car as one of the hottest objects; in practice, because white surfaces tend to reflect incident thermal infrared more than darker, more emissive materials, they should appear relatively cooler, and TherA is the only method that captures this behavior. Moreover, TherA is the only model that consistently identifies pedestrians with physically plausible heat-emitting characteristics. Since all evaluations are performed in a zero-shot setting on popular benchmark datasets, these results not only demonstrate TherA's strong generalization ability but also establish TherA as the first example-guided translation module explicitly specialized for RGB-to-TIR translation.

We do not fine-tune any of these baseline models on thermal data, so they fully benefit from large-scale RGB pre-training, yet the baseline reference-guided models still fail to recover physically meaningful thermal appearance. This demonstrates that RGB-to-TIR translation cannot be solved by generic appearance-based style transfer alone, and instead requires a thermal-aware model that explicitly accounts for semantic structure and imaging physics.

E. Limitations and Failure Cases

Dependence on RGB visibility and VLM conditioning. TherA generates pseudo-TIR images by conditioning a diffusion model on TherA-VLM. As a result, failure of either component can lead to suboptimal translations. We observe that errors are most common in frames with severely degraded RGB images, such as strong motion blur, almost completely dark scenes, or saturated/overexposed regions where scene geometry and texture are barely visible. In such cases TherA-VLM may produce incomplete or noisy descriptions and the diffusion model has little visual guid-

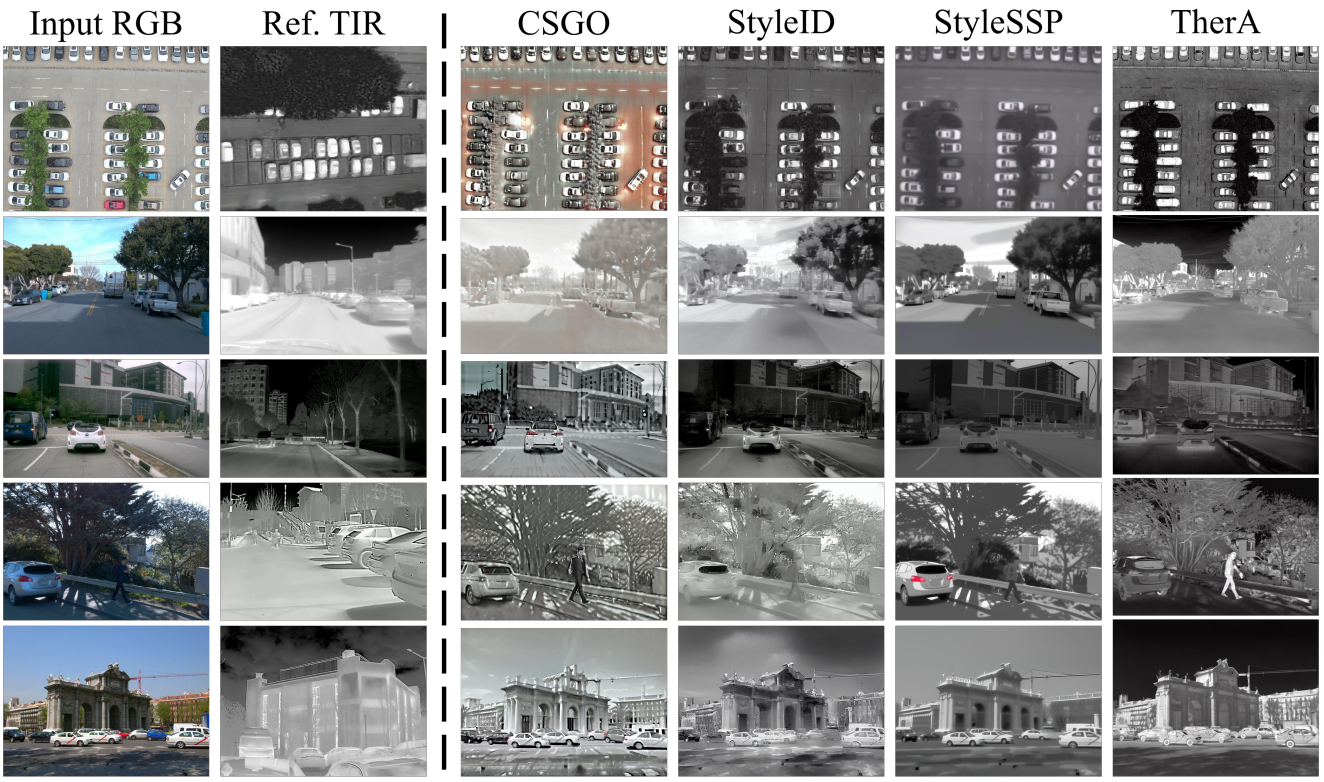


Figure 15. **Reference-guided translation comparison.** Qualitative comparison of existing reference-guided translation models and TherA. Source images taken from popular RGB benchmarks (from the top: Visdrone [71], Waymo [2], NuScenes [54], Waymo [2], and MegaDepth [55])

.

ance, which can result in unrealistic or unstable TIR predictions. Our prompting explicitly instructs the VLM to abstain from low-confidence statements, and user-side control can partially mitigate this by editing the instruction using text-guided or reference image-guided control, but the method fundamentally assumes that the RGB image contains sufficient visual information.

Relative temperature representation. Our current system operates on normalized thermal imagery: pixel intensities are normalized and only encode local temperature ordering and contrast, not absolute radiometric values. Consequently, TherA is not suitable for applications that require metric temperature estimation, emissivity calibration, or strict compliance with a full radiometric chain. Instead, our translations are designed to preserve qualitative thermal structure (relative hot/cold patterns, material- and scene-dependent contrast) and to support perception tasks such as recognition, matching, and segmentation. Extending the framework to calibrated radiometric data and explicitly modeling absolute temperature remains an important direction for future work.

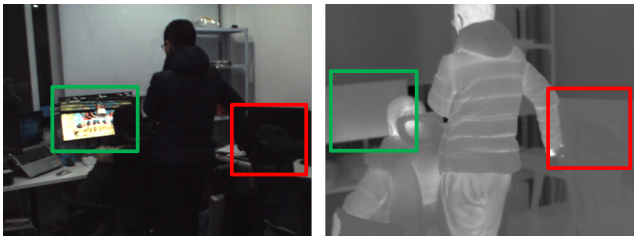


Figure 16. Limitation of controllable active/passive editing for indoor electronics. In the RGB image (left), a powered-on monitor (green box) and a powered-off monitor (red box) are clearly distinguishable. In the corresponding real TIR image (right), however, the temperature contrast between the two displays is subtle, making it difficult for TherA to produce a visually striking on/off effect for such objects.

Scope of active/passive controllability. While TherA can toggle the active/passive state of objects via reference image guidance, the visual effect is most pronounced for objects that exhibit strong and spatially extended heat emission, such as vehicles and human bodies in outdoor scenes. For other object categories, especially small indoor electronics (e.g., monitors, laptops) or surfaces with weak ther-

mal contrast, the change in the real TIR domain is often subtle. Figure 16 illustrates such a case: the RGB image contains both an active display and an inactive monitor, yet the corresponding thermal image shows only a modest intensity difference between them. Our translations reflect this limited contrast rather than a dramatic on/off switch, and indoor scenes with nearly uniform ambient temperature similarly offer limited room for controllable variation. This bias stems from the underlying training data distribution and highlights that our controllable attributes are currently most reliable for large-scale outdoor structures and vehicles, rather than all possible heat sources.

Looking ahead, we expect that stronger thermal-aware VLM backbones and joint end-to-end fine-tuning of the VLM and diffusion components on more diverse large-scale RGB–TIR data will further reduce these failure modes and improve robustness in challenging illumination regimes.