Rethinking Semantics for Complex-Scene Thermal Image Generation

Anonymous WACV Algorithms Track submission

Paper ID 5

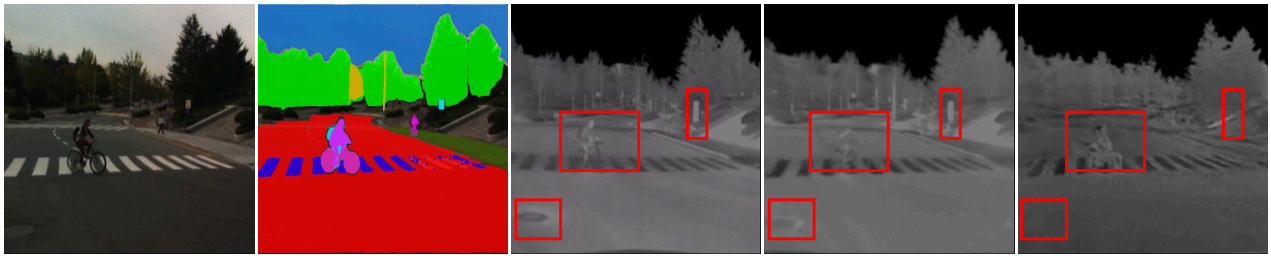

Figure 1. **The Impact of Emissivity Priors on Thermal Synthesis.** We investigate whether emissivity priors, provided as segmentation maps, enhance thermal image generation from RGB. Contrary to expectations, our analysis reveals that a baseline model (RGB2Thermal) [16] trained solely on RGB data produces high-fidelity thermal estimates consistent with the ground truth (GT). Introducing auxiliary semantic guidance (RGBSegMap2Thermal), however, disrupts the synthesis process, introducing artifacts and reducing fidelity, as highlighted in the red boxes. This counterintuitive result demonstrates that RGB context alone provides robust cues for thermal estimation, while segmentation priors can impair performance and model robustness.

Abstract

The scarcity of diverse thermal infrared datasets limits thermal vision research, motivating the generation of thermal images from RGB data. Prevailing insights assume extra contextual cues (e.g., segmentation maps) are essential priors for this task, yet this premise remains untested for complex scenes. We present a comprehensive empirical study that rigorously tests this assumption. Through extensive ablation, we reveal a counter-intuitive result that additional contextual information does not yield substantial gains. A model trained on RGB input alone not only matches but often surpasses the performance of more complex, context-aware architectures. This indicates that the visible spectrum is sufficiently rich for networks to learn latent thermal relationships. We supplement this finding with a novel analysis quantifying each RGB channel's contribution to the thermal output. Our results question the necessity of additional context in the form of segmentation maps and suggest a re-focusing of future work on effectively leveraging RGB data over integrating complex auxiliary inputs.

1. Introduction

The performance of computer vision systems is often limited by the operating principles of RGB cameras, which capture reflected light within the narrow visible spectrum $380\,\mathrm{nm}$ to $750\,\mathrm{nm}$. This dependence on ambient illumination and susceptibility to wavelength-dependent scattering renders them ineffective in low-light conditions or obscurant-filled environments like fog and smoke. In contrast, thermal infrared imaging ($8\,\mu\mathrm{m}$ to $14\,\mu\mathrm{m}$) detects infrared thermal radiation emitted directly by objects according to their temperature and emissivity. This fundamental shift from measuring reflected energy to emitted energy yields a modality that is both invariant to visible lighting and sensitive to material properties. As a result, thermal imaging is vital for safety-critical applications including autonomous navigation, surveillance, and medical diagnostics [3, 22, 24].

However, the development of robust thermal vision models is constrained by the scarcity of diverse, large-scale thermal infrared datasets, primarily due to high sensor costs and privacy concerns, which has spurred significant interest in cross-modal image synthesis, specifically generating realistic thermal imagery from readily available RGB inputs [6, 14]. The core challenge in translating RGB images

to thermal images arises from a fundamental phenomenon of physics. Materials that appear nearly identical in the visible spectrum can exhibit markedly different heat signatures in the thermal infrared domain. For example, a black-painted car body, asphalt road, and tinted glass may all appear similarly dark in an RGB image, yet they produce substantially different thermal signatures due to variations in emissivity, temperature, and reflectivity. This exemplifies what we define as *RGB–Thermal Mapping Ambiguity*: a many-to-many mapping problem where visually similar RGB regions can correspond to distinct thermal outputs, while visually distinct objects (e.g., differently colored cars) can yield similar thermal signatures. Moreover, the same object, such as a running car engine versus its body, can exhibit varying thermal appearances under different states. This inherent ambiguity violates the assumption of a deterministic cross-modal correspondence, undermining conventional supervised learning approaches that rely on pixel-level alignment.

A prevailing hypothesis to address this challenge suggests that augmenting RGB input with explicit contextual information, such as semantic segmentation or material maps, should provide critical emissivity priors to guide the synthesis of more physically plausible thermal outputs. This is particularly compelling given that thermal signatures are highly context-dependent, explicitly providing emissivity priors in the form of semantic maps should simplify the learning problem. Yet, despite its intuitive appeal, this concept remains largely unexplored and empirically unverified. Existing studies have been confined to overly simplistic, single-object scenarios while not evaluating on complex-image or multi-object scenarios [9, 21], leaving a significant research gap in understanding the value of semantic context for generating complex, multi-object thermal scenes remains entirely unknown.

This paper presents, to our knowledge, the first comprehensive study to rigorously investigate this hypothesis. We ask a critical question: Does incorporating additional contextual information (e.g., segmentation maps) actually improve the quality and fidelity of synthetic thermal imagery for complex scenes? To answer this, we conducted an extensive ablation study. We trained a series of models under identical conditions, systematically varying only the input modalities (RGB alone, RGB with model-predicted segmentation, RGB with ground-truth segmentation, and RGB with material maps) to isolate the effect of each contextual cue. Contrary to the prevailing assumption, our results demonstrate that additional contextual information does not yield better performance. This key insight, visualized in Figure 1, challenges a foundational assumption in multimodal learning for thermal synthesis. Quantitatively, our experiments confirm that a model trained solely on RGB input not only matches but also consistently surpasses more complex, context-aware architectures across established metrics. This suggests that the information within the RGB spectrum is itself sufficiently rich for the task; the network can implicitly learn the underlying relationships between color, texture, material, and thermal properties without explicit, emissivity guidance.

To further understand this result, we provide a novel analysis quantifying the contribution of each individual RGB color spectra to the generated thermal output, offering new insights into how deep learning models distill thermal cues from visible spectrum information.

Our primary contributions are:

1. A critical empirical study challenging the presumed necessity of additional contextual information (e.g., emissivity) for generating complex-scene thermal images, a previously unexplored hypothesis.
2. A rigorous ablation study demonstrating that a model trained on RGB data alone can achieve state-of-the-art performance, surpassing models augmented with semantic or material context.
3. A comprehensive analysis explaining this result, which reveals that the RGB modality inherently encodes sufficient latent information for accurate cross-spectral mapping, rendering explicit contextual inputs unnecessary.

2. Related Work

Thermal image synthesis has evolved along two main directions. Simulation-based methods model thermal emission through geometry, material emissivity, and environmental conditions, as demonstrated by [1, 13]. While physically grounded, these approaches often exhibit limited realism and scalability. Data-driven alternatives employ CNNs such as [7, 8], followed by GAN-based models including [9, 10, 27]. More recently, diffusion models have achieved state-of-the-art performance, with ThermalDiff [16] enabling high-fidelity RGB-conditioned thermal generation. Diffusion models also support rich conditioning mechanisms, and segmentation maps have proven highly effective in natural image synthesis [18, 23, 25]. This success has inspired the assumption that semantic maps might similarly enhance thermal generation, motivated by the material dependence of thermal properties. Preliminary attempts within GANs [9] and diffusion models [21] explore this idea, but evaluations remain restricted to narrow, object-centric scenarios, leaving open whether semantic context is beneficial in complex outdoor environments. Explainable AI methods provide tools for analyzing conditioning behavior in diffusion models. Grad-CAM [20] and diffusion-specific adaptations [15] enable visualization of structural and semantic focus, offering insights into why a model succeeds or fails under different conditioning strategies.

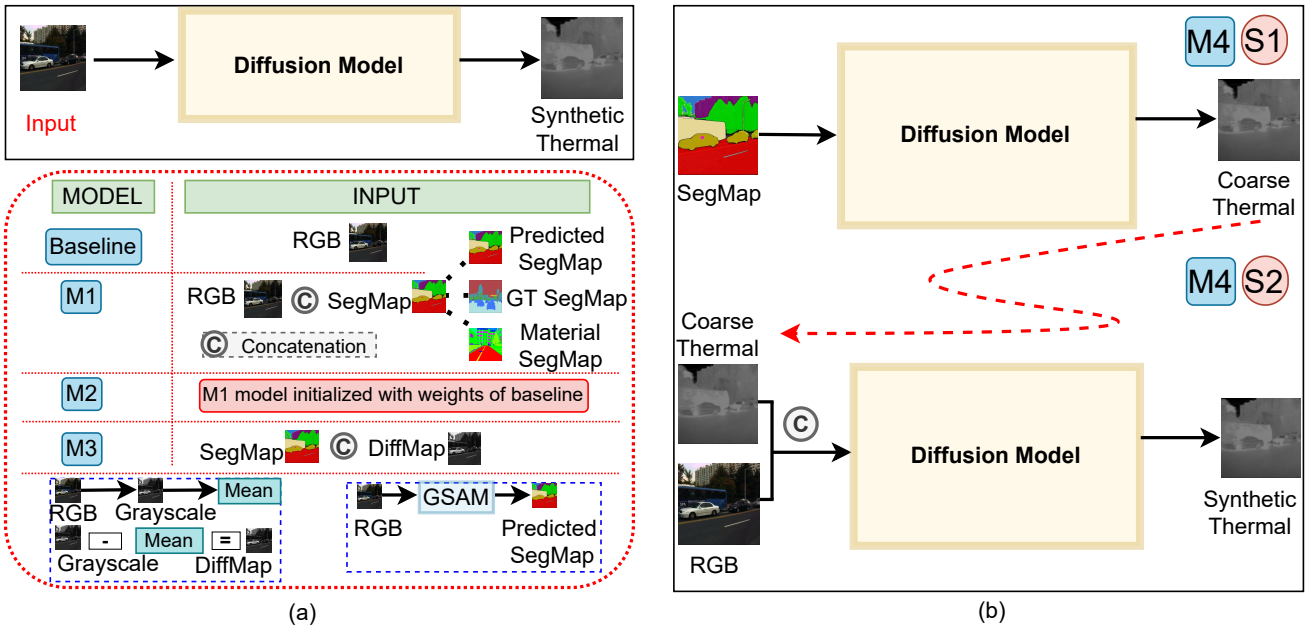

Figure 2. Overview of our methodology for investigating contextual conditioning for thermal image generation. **(a)** We build upon a baseline RGB-to-Thermal (Baseline) model [16] to evaluate if adding a segmentation map improves fidelity. Our experiments compare four conditioning strategies: **(M1) RGB+SegMap (Scratch):** Trained from scratch using three segmentation map variants: predicted maps from GSAM [17], ground truth maps, and material-specific maps. **(M2) RGB+SegMap:** M1 model initialized from the weights of pretrained baseline. **(M3) DiffMap+SegMap:** Uses a hand-crafted difference map derived from the RGB input, concatenated with a segmentation map. **(b) Two-Stage:** A cascaded approach where a coarse thermal image is first generated from the predicted segmentation map **(S1)**, which is then refined with RGB guidance **(S2)**.

3. Method

Our method builds on ThermalDiff [16], a conditional DDPM where a U-Net f_θ is trained to denoise a thermal image x_t conditioned on an RGB context $\mathbf{c}$ by minimizing $\mathcal{L}_c = \mathbb{E}[|\epsilon - f_\theta(x_t, \mathbf{c}, t)|^2]$, iteratively transforming noise into the final thermal estimate $\hat{x}_0$ during inference.

3.1. Architectural Variants for Context Investigation

To test our hypotheses, we develop four architectural variants of the ThermalDiff framework, fundamentally informed by the physics of thermal emission, to systematically evaluate how explicit material emissivity (ε) and implicit contextual cues influence thermal synthesis performance.

1) RGB and Segmentation Concatenation. This variant provides the model with both implicit cues from the visible spectrum and explicit material context. The RGB image V and the segmentation map S are concatenated channel-wise to form a unified 4-channel conditioning tensor $[V, S] \in \mathbb{R}^{4\times H\times W}$. The mapping is given by $\hat{I} = f_\theta(x_t, [V, S], t)$. This approach allows the network to learn the correlations between RGB texture/color (which hints at surface properties affecting ε) and the explicit material labels, leading to a more informed and physically-grounded synthesis.

2) Pre-trained Initialization. Identical in architecture to the above variant, this model is initialized with weights from a ThermalDiff model pre-trained for RGB-to-thermal synthesis ($\hat{I} = f_\theta(x_t, V, t)$) [16]. The additional segmentation channel in the input is initialized using Kaiming initialization [2]. This transfer learning approach tests whether leveraging robust feature representations learned from a large-scale RGB-thermal mapping task can accelerate convergence and enhance performance when augmented with explicit semantic context.

3) Thermal-Aware Difference Mapping. This physics-motivated variant incorporates a pre-processing step designed to highlight features correlated with thermal variation. The RGB image V is first converted to grayscale $G = 0.299R + 0.587G + 0.114B$. A thermal-aware difference map D is then computed by subtracting the spatial mean intensity: $D = G - \bar{G}$, where $\bar{G} = \frac{1}{HW}\sum_{i,j} G_{i,j}$. This operation amplifies local intensity variations, which often correlate with thermal gradients caused by surface orientation, micro-shading, and material transitions, factors that modulate local temperature T. The difference map D is then concatenated with the segmentation map S to form the conditioning input: $\hat{I} = f_\theta(x_t, [D, S], t)$. This design explicitly guides the model to focus on local contrast features

that are physically indicative of thermal patterns, providing a strong inductive bias.

4) Two-Stage Coarse-to-Fine Synthesis. This hierarchical variant decouples the two primary factors governing thermal emission—material emissivity (ε) and temperature (T). The approach reflects the underlying physics wherein the global thermal structure is largely determined by material type (captured by segmentation), whereas fine-scale variations depend on local temperature changes influenced by geometry and scene context (captured by RGB texture). In the first stage, a diffusion model generates a coarse thermal estimate $\hat{I}$coarse using only the segmentation map, establishing material-driven emissivity patterns: $\hat{I}\text{coarse} = g_\phi(x_t^{(1)}, S, t)$. The second stage refines this estimate by conditioning on both the RGB image V and the coarse prediction, producing a high-fidelity output $\hat{I}\text{fine} = h\psi(x_t^{(2)}, [V, \hat{I}_{\text{coarse}}], t)$. This cascaded formulation enables a physics-informed coarse-to-fine decomposition, resolving emissivity first and then refining temperature-dependent details.

4. Experiments

4.1. Datasets, Model and Metrics

Our primary evaluation uses the KAIST multispectral dataset [5], a standard urban benchmark, with 12,538/2,252 image pairs for training/testing. Semantic maps for conditioning are generated from RGB using Grounded-SAM [17]. For ground-truth ablations, we supplement with the FMB dataset for segmentation [12] and the MCubeS dataset for material analysis [11]. MCubeS provides per-pixel material maps across 20 classes; due to its limited size, models are fine-tuned on it after pre-training on KAIST. All training and evaluation are performed at 512×512 resolution using official dataset splits. We base all experiments on the ThermalDiff architecture [16], the current state-of-the-art diffusion model for RGB-to-thermal synthesis. Using this fixed, high-performing backbone isolates the effect of conditioning inputs from architectural improvements. We employ a standard suite of metrics: PSNR and L1 for pixel-wise accuracy; SSIM and LPIPS [26] for structural/perceptual similarity; and FID [4] for distributional fidelity.

4.2. State-of-the-art comparison for thermal image synthesis

4.2.1. Generation with predicted segmentation map

On the KAIST dataset [5], explicit semantic conditioning consistently degrades performance with the RGB-only baseline achieving superior results Table 1. Even sophisticated conditioning variants, including pretrained fine-tuning, hand-engineered priors, and two-stage generation, fail to match the baseline's reconstruction quality or collapse in diversity, indicating that semantic maps disrupt rather than enhance the synthesis process.

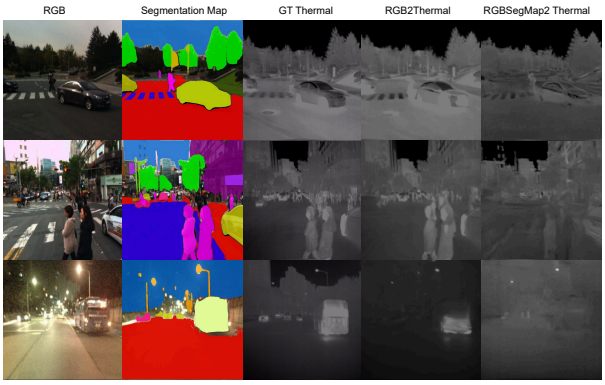

Figure 3. Qualitative comparison of thermal image synthesis on the KAIST dataset [5].The RGB2Th model produces sharper, more physically consistent reconstructions, while RGB+SegMap2Th shows reduced structural and temperature fidelity despite additional semantic guidance, aligning with quantitative results in Table 1. (Note: Brightness and contrast of RGB+SegMap2Th samples are enhanced for visualization.)

Table 1. Performance comparison on the KAIST dataset [5] using different contextual conditioning strategies. All models are trained from scratch using semantic maps from GSAM [17] and our computed difference map (DiffMap, see Sec. 3). Model 4 additionally utilizes coarse thermal guidance (thermal images generated from segmentation maps alone, as detailed in Sec. 3).

Model	RGB	SegMap	DiffMap	Coarse Th.	PSNR↑	SSIM↑	L1↓	LPIPS↓	FID↓
KAIST [5]									
Baseline [16]	✓	✗	✗	✗	**30.29**	**0.85**	**0.02**	**0.11**	23.09
Model 1	✓	✓	✗	✗	23.44	0.64	0.06	0.26	29.33
Model 2	✓	✓	✗	✗	25.67	0.71	0.20	0.20	**16.77**
Model 3	✗	✓	✓	✗	22.86	0.62	0.07	0.27	26.39
Model 4 (S1)	✗	✓	✗	✗	21.22	0.57	0.08	0.33	24.58
Model 4 (S2)	✓	✗	✗	✓	19.43	0.52	0.10	0.42	77.77

4.2.2. Generation with ground-truth segmentation map

Using the FMB dataset with ground-truth annotations (Table 2), Baseline (RGB-only) again outperforms Model 1 (RGB + GT SegMap) across all metrics. This confirms that semantic conditioning limitations persist even with perfect annotations, highlighting the redundancy of explicit semantic guidance.

4.2.3. Impact of material-specific segmentation map

On the MCubeS dataset (Table 3), material-specific segmentation yields a nuanced outcome: Model 1 marginally improves FID but severely compromises reconstruction accuracy. This underscores a critical trade-off between statistical alignment and per-image fidelity.

Table 2. Performance comparison on the FMB dataset [12]. The baseline model uses RGB context only, while Model 1 is trained from scratch using ground-truth object segmentation annotations in addition to RGB input.

Model	RGB	SegMap	DiffMap	Coarse Th.	PSNR↑	SSIM↑	L1↓	LPIPS↓	FID↓
FMB [12]									
Baseline [16]	✓	✗	✗	✗	**21.34**	**0.72**	**0.07**	**0.31**	**40.23**
Model 1	✓	✓	✗	✗	18.71	0.67	0.10	0.38	45.40

Table 3. Performance comparison on the MCubeS dataset [11]. The baseline model uses RGB context only, while Model 1 is initialized with weights pre-trained on KAIST and subsequently fine-tuned using ground-truth material segmentation maps.

Model	RGB	SegMap	DiffMap	Coarse Th.	PSNR↑	SSIM↑	L1↓	LPIPS↓	FID↓
MCubeS [11]									
Baseline [16]	✓	✗	✗	✗	**42.80**	**0.92**	**0.0046**	**0.09**	235.75
Model 1	✓	✓	✗	✗	25.52	0.72	0.05	0.19	**225.31**

Table 4. Contribution of individual RGB channels to reconstruction metrics on the KAIST dataset [5].

Channels	PSNR↑	SSIM↑	L1↓	LPIPS↓	FID↓
Red Channel	22.75	0.61	0.0771	0.278	32.53
Green Channel	23.17	0.63	0.0701	0.271	32.86
Blue Channel	23.15	0.61	0.0702	0.270	32.35

4.3. Investigating the Spectral Contribution of RGB Channels

We analyze the spectral contribution of RGB channels using the KAIST dataset [5], with each channel providing distinct yet complementary cues for thermal prediction. The red channel (≈620–750 nm) captures superficial material properties but lacks structural detail; the green channel (≈495–570 nm) encodes strong structural and textural information, achieving the highest PSNR and SSIM; and the blue channel (≈450–495 nm) provides high-frequency edge detail but is susceptible to scattering, yielding the best LPIPS and FID. Only their combined fusion bridges the spectral gap to produce accurate thermal reconstructions, confirming RGB as a strong contextual prior. Statistical tests comparing distributions and supporting visualizations are provided in the supplementary material S4, S1, S2.

4.4. Understanding thermal image generation via visual analysis

We analyze denoising trajectories, internal feature prioritization, and timestep-specific concept learning in both our baseline [16] and segmentation-conditioned model (Model 1). The baseline exhibits coherent, hierarchical refinement, recovering coarse structures by mid-trajectory, establishing global layout in early timesteps, and refining local details in late steps. Diffusion Gradient-weighted Class Activation Mapping (DF-CAM) [15] confirms it learns a balanced representation, with down-sampling layers prioritizing global structure, mid-layers building coherence, and up-sampling layers refining details. In contrast, the segmentation-conditioned model displays unstable recovery, overfits to segmentation boundaries in its activations at the expense of holistic context, and lacks coherent timestep specialization, resulting in fragmented internal representations and discordant output. This absence of hierarchical learning confirms that the semantic prior acts as a disruptive noise source rather than a useful guide. Visualizations are provided in the supplementary material S3, S4, S5.

5. Discussion and Conclusion

Our results demonstrate that explicit semantic conditioning does not improve thermal image synthesis and often degrades performance relative to RGB-only baselines. This overturns the assumption that segmentation maps act as meaningful physical priors. RGB inputs already encode sufficient structural, textural, and material cues for inferring thermal properties, whereas semantic maps introduce noise that destabilises the generative process. The degradation persists even with ground-truth segmentation and material labels, indicating that conventional annotations lack emissivity-related information required for thermal reconstruction. Moreover, semantic categories provide only coarse abstractions of heterogeneous materials—such as skin, hair, fabric, metal, and glass—creating ambiguity rather than offering useful guidance.

To isolate the effect of semantic priors, we use an in-domain, state-of-the-art diffusion backbone [16] with a standard concatenation-based conditioning strategy [19]. The consistent failure of this controlled setup to yield gains confirms that the limitation lies in the semantic prior itself, not in architectural or fusion-mechanism choices.

Evaluating thermal synthesis is also challenging, as a single RGB scene can correspond to multiple valid thermal outputs governed by unobservable factors such as emissivity, kinetic temperature, and ambient reflections. While we report standard reconstruction and perceptual metrics, we further validate our findings through downstream object detection experiments on generated thermal imagery (results to appear in the camera-ready version).

Overall, this study shows that RGB-only context is superior for thermal image synthesis, and semantic maps do not enhance output quality. Future work should therefore focus on strengthening core RGB-to-thermal mapping rather than incorporating redundant semantic priors, reframing research directions for high-fidelity thermal image generation.

References

[1] R. Guissin, E. Lavi, A. Palatnik, Y. Gronau, E. Repasi, W. Wittenstein, R. Gal, and M. Ben-Ezra. Irisim: Infrared imaging simulator. In *SPIE Defense + Commercial Sensing*, 2005. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 3

[3] 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. 1

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4

[5] 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*, pages 1037–1045, 2015. 4, 5, 9

[6] Yumi Iwashita, Kazuto Nakashima, Sir Rafol, Adrian Stoica, and Ryo Kurazume. Mu-net: Deep learning-based thermal ir image estimation from rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[7] 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. 2

[8] 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. 2

[9] Vladimir V Kniaz, Vladimir A Knyaz, Jiri Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2, 8

[10] L. Li, P. Li, M. Yang, and S. Gao. Multi-branch semantic gan for infrared image generation from optical image. In *Intelligence Science and Big Data Engineering*, pages 484–494. Springer, 2019. 2

[11] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19800–19808, 2022. 4, 5

[12] 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*, pages 8115–8124, 2023. 4, 5

[13] Neelu Madan, Mia Sandra Nicole Siemon, Magnus Kaufmann Gjerde, Bastian Starup Petersson, Arijus Grotuzas, Malthe Aaholm Esbensen, Ivan Adriyanov Nikolov, Mark Philip Philipsen, Kamal Nasrollahi, and Thomas B Moeslund. Thermalsynth: A novel approach for generating synthetic thermal human scenarios. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 130–139, 2023. 2

[14] Mehmet Akif Özkanoğlu and Sedat Ozer. Infragan: A gan architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155:69–76, 2022. 1

[15] Ji-Hoon Park, Yeong-Joon Ju, and Seong-Whan Lee. Explaining generative diffusion models via visual analysis for interpretable decision-making process. *Expert Systems with Applications*, 248:123231, 2024. 2, 5, 9, 10

[16] 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. 1, 2, 3, 4, 5, 8, 9, 10, 11

[17] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3, 4

[18] 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*, pages 10684–10695, 2022. 2

[19] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 5

[20] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[21] Doan Thinh Vo, Phan Anh c, Nguyen Nhu Thao, and Huong Ninh. An approach to synthesize thermal infrared ship images. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024. 2, 8

[22] Michael Vollmer and Klaus-Peter Möllmann. *Infrared thermal imaging: fundamentals, research and applications*. John Wiley & Sons, 2018. 1

[23] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 2

[24] 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. 1

[25] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In

Proceedings of the IEEE/CVF international conference on computer vision, pages 3836–3847, 2023. 2

[26] 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*, pages 586–595, 2018. 4

[27] M. A. Özkanoglu and S. Ozer. Infragan: A gan architecture to transfer visible images to infrared domain. *Pattern Recognition Letters*, 155:69–76, 2022. 2

Supplementary Material

S1. Problem Definition

While RGB-to-thermal image synthesis has been extensively studied, a critical question remains not only unresolved but largely unexamined, whether incorporating additional contextual information, such as semantic maps, genuinely enhance thermal synthesis quality for complex, multi-object scenes, or is its benefit overstated and confined to simplistic benchmarks? A prevalent hypothesis posits that augmenting RGB input with contextual information, such as semantic segmentation map should provide critical emissivity priors to guide the synthesis of more physically plausible thermal outputs. The intuition is that thermal signatures are highly material-dependent, and explicitly providing this context should simplify the learning problem.

However, this widely-accepted assumption remains largely empirically unverified for complex, multi-object scenarios. Existing studies [9, 21] validating the benefits of semantic context have been confined to overly simplistic, single-object cases. The core objective of this work is to rigorously investigate two fundamental questions:

1. **Context Effectiveness**: Does incorporating additional contextual information (such as semantic segmentation maps) actually improve thermal image synthesis fidelity and quality for complex, multi-object scenes compared to RGB-only approaches?
2. **RGB Channel Analysis**: What is the individual contribution of each RGB channel (Red, Green, Blue) towards thermal generation, and how do these spectral components inform the cross-modal mapping process?

S2. Mapping Paradigms

Formally, given a dataset of paired images $\mathcal{D} = (V_i, I_i)_{i=1}^N$, where $V_i \in \mathbb{R}^{3\times H\times W}$ is an RGB image and $I_i \in \mathbb{R}^{H\times W}$ is its corresponding thermal counterpart, we investigate two competing mapping paradigms:

RGB-Only Mapping: $\hat{I} = f_\phi(V)$, where $V \in \mathbb{R}^{3\times H\times W}$ is the input RGB image and ϕ denotes the learnable parameters of the mapping function. The use of visible spectrum images as contextual guidance for thermal synthesis is well-established in modern computer vision approaches. Building on pioneering work by ThermalDiff [16], we recognize that RGB imagery provides rich spectral information that correlates strongly with thermal emission properties. Specifically, different surface materials exhibit characteristic reflectance profiles across the red, green, and blue spectral bands that correspond to their thermal emission characteristics. This relationship enables RGB context to provide critical cues related to surface geometry, material composition, scene structure, and environmental conditions. **Context-Augmented Mapping:** $\hat{I} = g_\psi(V, S)$, where $S \in \mathbb{R}^{1\times H\times W}$ represents a single-channel semantic map with spatial dimensions $H \times W$ and ψ denotes the learnable parameters of the mapping function g. While RGB images provide implicit material cues through spectral signatures, semantic segmentation maps offer explicit, pixel-wise material identification that directly informs the underlying physics of thermal emission. The fundamental challenge in thermal synthesis lies in accurately predicting material-specific emissivity values ε, which serve as the primary scaling factor for total thermal radiance [1] according to the Stefan-Boltzmann law: $W = \varepsilon\sigma T^4$ where W represents the total power radiated per unit area, T is the absolute temperature (Kelvin) and σ is the Stefan-Boltzmann constant.

Semantic segmentation maps establish a direct mapping from spatial coordinates (u, v) to material class $c \in 1, ..., C$, where each class corresponds to a characteristic emissivity value ε_c: $\varepsilon(u, v) = \mathcal{E}(S(u, v))$ Here, $\mathcal{E} : c \to \varepsilon_c$ represents a known physical mapping from semantic categories to material emissivity properties (e.g., $\mathcal{E}$('human skin') ≈ 0.98, $\mathcal{E}$('polished metal') ≈ 0.05).

This explicit material prior addresses fundamental ambiguities that challenge RGB-only approaches. For instance, while dark-colored automotive surfaces (low ε) and dark asphalt paving (high ε) may exhibit nearly identical RGB appearances, they possess radically different thermal emission characteristics due to their material composition. The semantic segmentation map resolves this ambiguity by providing categorical labels ('vehicle' vs. 'road'), thereby enabling physics-informed thermal prediction that incorporates known emissivity priors. This explicit material guidance theoretically enhances synthesis accuracy for complex multi-object scenes where precise material boundaries are critical for accurate temperature field estimation.

S3. Investigating the Spectral Contribution of RGB Channels

To quantify the individual contribution of each spectral band to thermal synthesis, we decompose the RGB input into its constituent channels: $V = [V_R, V_G, V_B]$, where $V_c \in \mathbb{R}^{1\times H\times W}$ for $c \in R, G, B$

This decomposition enables targeted analysis of how different wavelengths inform thermal prediction. Each channel captures distinct material properties: red spectra correlate with surface composition and thermal proxies, green bands associate with vegetation and moisture content, and blue channels relate to surface reflectivity and material type. These spectral signatures provide complementary cues for

[1] It is important to note that thermal cameras in the LWIR band measure total infrared radiance, which includes both emitted radiation (function of temperature and emissivity) and reflected radiation from the environment. While emissivity is the primary material property affecting thermal emission, reflectivity also plays a role in determining the final thermal signature observed by the camera.

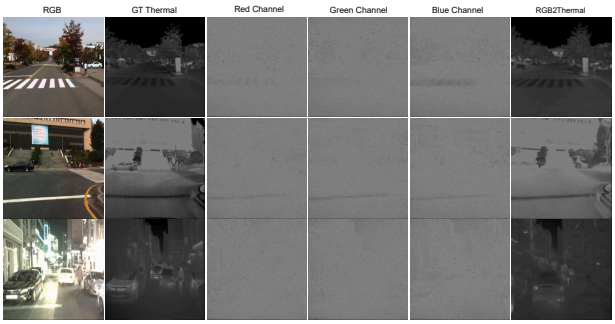

Figure S1. Contribution analysis of Red, Green, and Blue channels for thermal synthesis on the KAIST LWIR dataset [5]. As observed, single channels provide partial cues, red captures coarse reflectance, green encodes sharper textures, and blue enhances edges but only their combination yields synthesized thermal images closely matching ground truth.

inferring emissivity and thermal behavior.

We analyze the role of individual RGB channels in thermal synthesis using the KAIST Multispectral Dataset [5]. As quantified in Table 4 and visualized in Figure S1, each channel provides distinct yet complementary information for thermal prediction.

Red Channel. Operating at the longest visible wavelengths (≈620–750 nm), this channel captures superficial material properties and provides weak thermal priors related to emissivity, but lacks critical structural detail when used in isolation.

Green Channel. Encoding the strongest structural and textural information (≈495–570 nm), this channel offers valuable spatial cues that correlate highly with thermal scene layout, achieving the highest PSNR (23.17 dB) and SSIM (0.63) among individual channels.

Blue Channel. Sensitive to high-frequency edges and contrasts (≈450–495 nm), this channel provides supplementary detail but is highly susceptible to atmospheric scattering, achieving the best LPIPS (0.270) and FID (32.35) scores.

As clearly visualized in Figure S1, single channels provide partial cues, red captures coarse reflectance, green encodes sharper textures, and blue enhances edges but only their combination yields synthesized thermal images closely matching ground truth. This complementary fusion bridges the spectral gap, confirming RGB imagery as a strong contextual prior for thermal synthesis.

S4. Statistical Significance Testing

To quantitatively assess the realism of the generated thermal images, we compare the distributions of ground truth (GT) thermal data against those produced by the RGB2Thermal baseline [16] and our RGB+SegMap model (Model 1) using violin plots (Figure S2). The distributions are evaluated using three statistical measures: mean intensity, standard deviation, and kurtosis.

Our analysis reveals that the RGB2Thermal baseline produces a mean intensity slightly higher than the GT, indicating a mild but consistent over-estimation bias. In contrast, the RGB+SegMap model significantly under-estimates the mean, suggesting a substantial loss of thermal fidelity. In terms of variance, the baseline model maintains a standard deviation closely aligned with the GT, demonstrating stable generative diversity. The RGB+SegMap model, however, suffers from a severe collapse in variance, resulting in overly homogeneous and distributionally poor outputs.

Regarding kurtosis, the baseline yields moderately heavy-tailed distributions, retaining a degree of fine-scale variability. The RGB+SegMap model also produces heavy-tailed distributions but exhibits less stable peakedness, implying that semantic conditioning introduces training instabilities and leads to unreliable distributional alignment.

The training progression of the RGB+SegMap model further highlights its instability: it begins with extreme over-estimation and high dispersion, briefly approaches GT alignment mid-training, but ultimately diverges toward under-estimation and variance collapse. This erratic behavior indicates that semantic conditioning, while potentially informative, adversely affects convergence stability compared to the RGB-only baseline.

S5. Understanding thermal image generation via visual analysis

In Figure S3, Figure S4 and Figure S5, we present visual comparison of the denoising trajectories for the RGB2Th baseline [16] and the RGB+SegMap2Th model (Model 1).

S5.1. Regions recovered by the diffusion model in terms of semantic and detail level

We analyze the denoising trajectories of our baseline (RGB2Th) and segmentation-conditioned model (RGB+SegMap2Th). The baseline exhibits coherent, progressive refinement: coarse scene structures (e.g., vehicle contours, road layout) emerge by mid-denoising steps ($t \approx 1000$), with finer details consolidating in the final steps ($t < 200$). In contrast, the segmentation-conditioned model delays semantic recovery, suppressing object boundaries until later stages. This results in intermediate instability as the model struggles to reconcile the noisy segmentation prior with latent scene structure. While both models converge to plausible outputs, the baseline achieves smoother semantic alignment.

S5.2. Visual concepts prioritized by a given context

We employ Diffusion Gradient-weighted Class Activation Mapping (DF-CAM) [15] to visualize internal feature representations. The RGB2Thermal model [16] demonstrates a

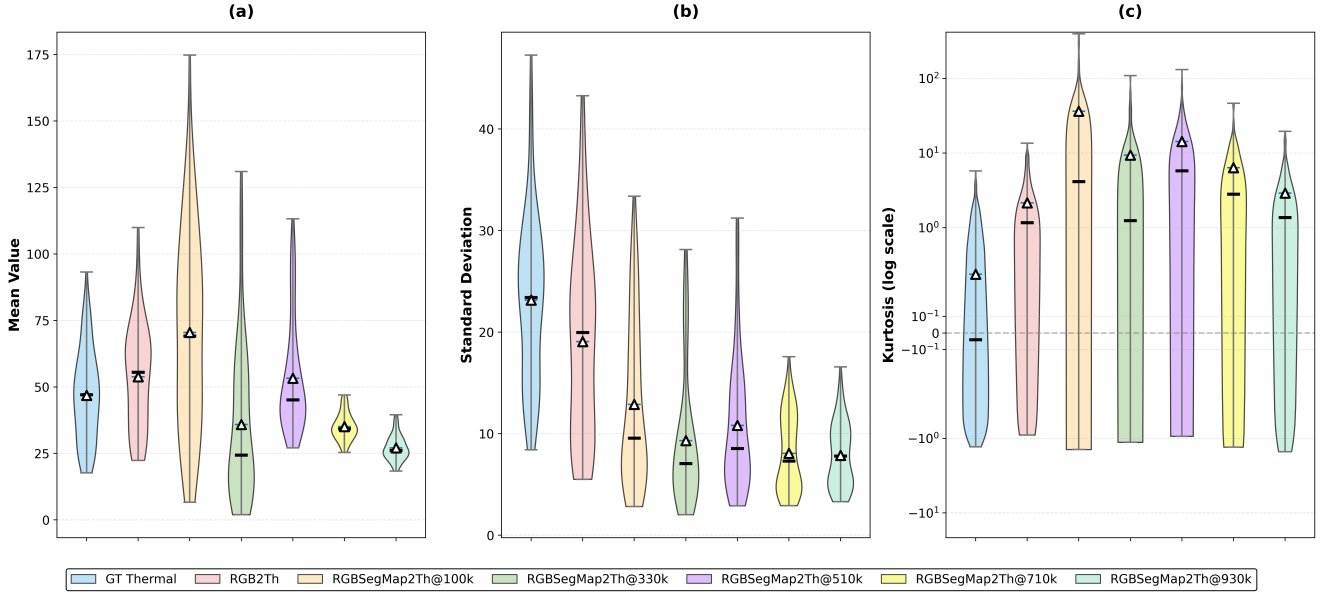

Figure S2. Statistical comparison of thermal image distributions: ground truth (GT), RGB2Th baseline [16], and RGB+SegMap2Th (Model 1) at various training stages. Violin plots show distributions of mean intensity, standard deviation, and kurtosis. RGB2Th maintains consistent but biased statistics, while RGB+SegMap2Th shows optimal alignment at intermediate training (510k) but deteriorates with extended training (930k).

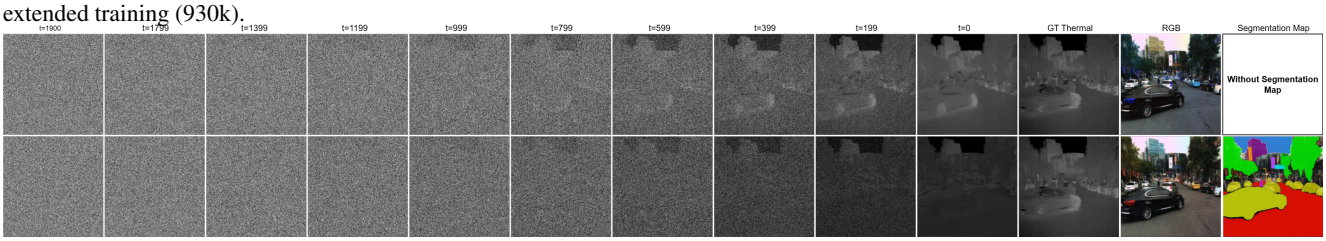

Figure S3. Visual comparison of the denoising trajectories for the RGB2Th baseline [16] (top) and the RGB+SegMap2Th model, Model 1 (bottom). The process evolves from noise ($t = 2000$) to the final generated image ($t = 0$). The rightmost column shows the conditioning input for each row: RGB image for the baseline, and both the RGB and segmentation map for Model 1. The baseline model shows more coherent and progressive semantic structuring compared to Model 1, which exhibits delayed and less stable recovery due to its noisy segmentation prior. (Note: Brightness and contrast of RGB+SegMap2Th samples are enhanced for visualization).

hierarchical focus: down-sampling layers emphasize global structural cues like road boundaries and vehicle silhouettes; mid-layers build semantic and spatial coherence; and up-sampling layers refine fine-grained details.

Conversely, the RGB+SegMap2Th model shows distorted prioritization. Its down- and mid-layer activations overfit to segmentation boundaries, focusing on noisy semantic regions at the expense of holistic contextual understanding. This impairs detail refinement in up-sampling layers, resulting in fragmented activations and diminished reconstruction quality. These findings confirm that while the baseline learns a balanced, hierarchical representation, the segmentation-conditioned model's over-reliance on its semantic prior disrupts this process, leading to reduced consistency and inferior output.

S5.3. Visual concept implied at timestep t

Analysis using non-uniform sampling [15] reveals fundamental differences in timestep usage. The RGB2Th baseline [16] exhibits a structured process: early timesteps ($t > 1500$) capture global semantics and thermal layout, while late timesteps ($t < 400$) refine local textures and contrast.

In contrast, the RGB+SegMap2Th model shows no coherent learning phase specialization. Early sampling fails to establish a robust semantic layout due to the noisy segmentation prior, while late sampling generates discordant local details decoupled from global structure. This absence of hierarchical learning demonstrates that the semantic prior provides no useful information, instead acting as a disruptive source of noise.

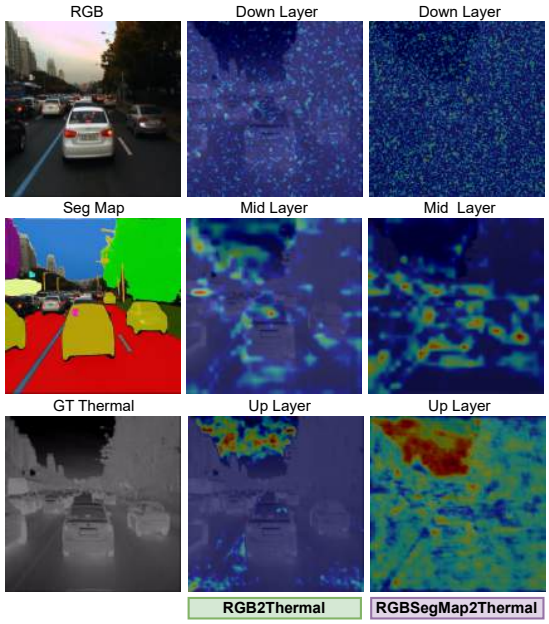

Figure S4. Layer-wise DF-CAM (Diffusion Gradient-weighted Class Activation Mapping) visualizations of U-Net feature activations. The visualizations show contextual inputs (RGB or RGB+SegMap) with corresponding ground truth thermal imagery. Activations from the down, mid, and up layers of the RGB2Th baseline [16] model emphasize global structural cues in early layers and progressively refine semantic and thermal details in later stages. In contrast, activations from the RGB+SegMap2Th model display over-sensitivity to segmentation boundaries and fragmented refinement, highlighting the disruptive influence of noisy semantic priors on thermal image generation.

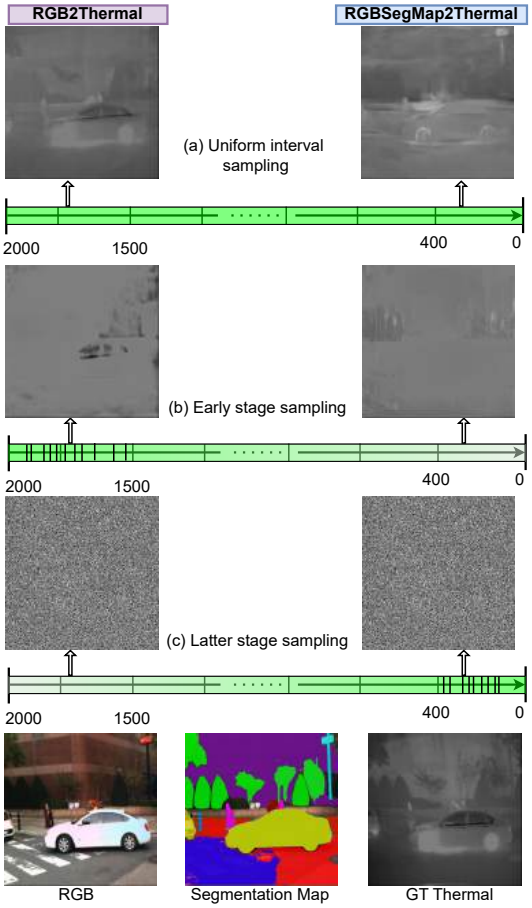

Figure S5. Effect of timestep sampling strategy on generation quality. The RGB2Th baseline [16] (left) maintains a coherent structure across all strategies. The RGB+SegMap2Th model (right) is highly unstable: early sampling ($t > 1500$) fractures semantics by overfitting to the segmentation map, while late sampling ($t < 400$) produces incoherent local details. The results confirm that the additional segmentation context does not enhance generated image fidelity and introduces a critical dependency on balanced (uniform) training for stability. (Note: Brightness and contrast of RGB+SegMap2Th samples are enhanced for visualization.)