Inference-Time Scaling of Diffusion Models for Infrared Data Generation

Kai A. Horstmann[*]
Cornell University

Maxim Clouser
YRIKKA, Inc.

Kia Khezeli
YRIKKA, Inc.

Abstract

Infrared imagery enables temperature-based scene understanding using passive sensors, particularly under conditions of low visibility where traditional RGB imaging fails. Yet, developing downstream vision models for infrared applications is hindered by the scarcity of high-quality annotated data, due to the specialized expertise required for infrared annotation. While synthetic infrared image generation has the potential to accelerate model development by providing large-scale, diverse training data, training foundation-level generative diffusion models in the infrared domain has remained elusive due to limited datasets. In light of such data constraints, we explore an inference-time scaling approach using a domain-adapted CLIP-based verifier for enhanced infrared image generation quality. We adapt FLUX.1-dev, a state-of-the-art text-to-image diffusion model, to the infrared domain by finetuning it on a small sample of infrared images using parameter-efficient techniques. The trained verifier is then employed during inference to guide the diffusion sampling process toward higher quality infrared generations that better align with input text prompts. Empirically, we find that our approach leads to consistent improvements in generation quality, reducing FID scores on the KAIST Multispectral Pedestrian Detection Benchmark dataset by 10% compared to unguided baseline samples. Our results suggest that inference-time guidance offers a promising direction for bridging the domain gap in low-data infrared settings.

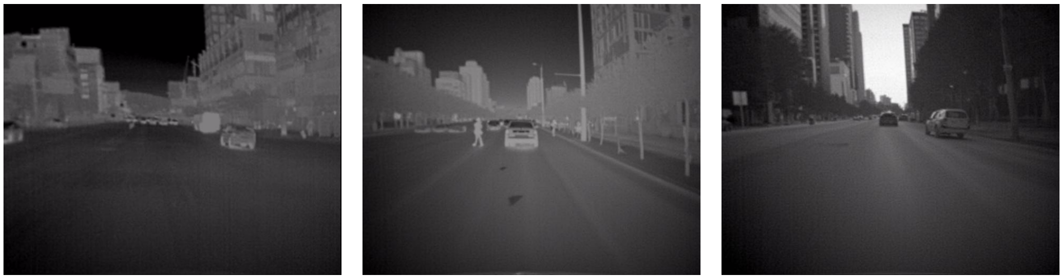
1 Introduction

Infrared sensors capture thermal information unavailable through conventional RGB cameras, making them essential for applications ranging from autonomous driving to medical imaging. The growing deployment of computer vision systems in such domains has created demand for large-scale synthetic infrared datasets to train robust downstream models such as object detectors and scene classifiers. Diffusion models [1, 2] have emerged as a promising approach for generating such synthetic data, offering the potential to create diverse, high-quality images that can augment limited real-world datasets.

While the ubiquity of color images allows large-scale diffusion models to be pretrained on billions of RGB images sourced from the web, infrared images remain far less accessible and are typically confined to a limited number of curated public datasets. This scarcity poses significant challenges for training foundation-level infrared diffusion models. Instead, finetuning pretrained RGB models leverages the rich world knowledge encoded from billions of natural images while adapting representations to the infrared modality. To this end, prior works [3, 4] have explored techniques to finetune pretrained RGB diffusion models on available public infrared image datasets, albeit with limited physical realism. Such approaches suffer from a lack of contextual grounding; while generated outputs may appear thermodynamically plausible, they often fail to remain consistent with the semantic or contextual cues provided by the conditioning input, whether a textual prompt or RGB

[*] Correspondence: kah288@cornell.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning to Sense.



IRScore (×10) (ours):	0.5062
IR Similarity:	0.4879
Grayscale Similarity:	0.3866
IRScore (×10) (pretrained):	0.1016

(a) Ground Truth IR

IRScore (×10) (ours):	0.5135
IR Similarity:	0.4213
Grayscale Similarity:	0.3186
IRScore (×10) (pretrained):	-0.0414

(b) High-Scoring Synthetic IR

IRScore (×10) (ours):	-0.3212
IR Similarity:	0.3189
Grayscale Similarity:	0.3831
IRScore (×10) (pretrained):	-0.0125

(c) Low-Scoring Synthetic IR

Figure 1: Example verifier scores for a ground truth infrared image and two synthetic images (generated from different random seeds) corresponding to the caption, "`A city street at dusk features tall buildings with illuminated signs, a marked road with directional arrows, and vehicles including a white SUV driving away from the camera.`" IRScore (ours), IR Similarity, and Grayscale Similarity are computed using our finetuned CLIP model, while IRScore (pretrained) relies on a pretrained CLIP model. IRScore, IR Similarity, and Grayscale Similarity correspond to Equation 1, and its unscaled first and second terms, respectively.

image. Other approaches such as PID [5] train diffusion models directly on limited infrared data with physics-informed constraints, but exhibit inconsistent generation quality and mode collapse, highlighting the pitfalls of direct training in the absence of large-scale pretraining.

To address such challenges, we turn to inference-time scaling, an extension of training-time neural scaling laws [6] commonly observed during the training of deep learning models. Inference-time scaling methods aim to improve the performance of models post hoc by expending additional compute during inference, and have shown promise in their recent application to large language models [7, 8, 9] and diffusion models [10, 11, 12]. Motivated by these advances, we investigate the efficacy of inference-time scaling in improving the quality of synthetic infrared images generated by diffusion models, a direction which, to the best of our knowledge, has not previously been explored. Whereas large-scale pretraining is constrained by infrared data scarcity, we now shift the computational burden from training to generation, thereby improving the visual quality of infrared images produced by diffusion models finetuned on limited data. Building on the inference-time scaling framework of Ma et al. [10] which combines sampling algorithms with verifiers, we extend this approach to the infrared domain and introduce a novel self-supervised verifier designed to evaluate the consistency and realism of synthetic infrared images. In particular, we finetune CLIP [13] to distinguish true infrared images from their grayscale counterparts, and leverage this model as an inference-time verifier to guide iterative search through the noise space, selecting noise latents that yield the highest-quality generations. Our experiments demonstrate measurable improvements in FID scores, and we further assess how different noise search strategies and verifier training protocols influence inference-time scaling performance.

2 Background

A naive approach to scaling diffusion models at inference time is to increase the number of denoising steps during image generation. Scaling up denoising steps can improve the visual quality of generated images, albeit with diminishing returns [14]. Recent works [10, 11, 12] have investigated alternative methods to scaling, motivated by the observation that different noise latents result in varied levels of quality in generated images [15, 16]. Ma et al. [10] extends this finding by introducing a scaling approach consisting of two components: verifiers and search algorithms.

2.1 Verifiers

Concretely, a verifier is defined as a function $\mathcal{V} : \mathbb{R}^{h \times w} \times \mathbb{R}^d \to \mathbb{R}$ which takes in a sample image and an optional condition (e.g., a prompt or image class) and produces a scalar value reflecting the quality of the image, either unconditionally or relative to the provided condition. In the text-to-image setting, Ma et al. [10] explores numerous pretrained models as verifiers, each designed to evaluate different notions of image quality. CLIPScore [17] is chosen for measuring image-prompt alignment, Aesthetic Score Predictor [18] for human aesthetic preferences, and ImageReward [19] for a more comprehensive metric incorporating both aesthetic quality and prompt adherence. Additionally, vision language models (VLMs) such as Gemini-1.5 Flash are employed for their sophisticated text-image understanding and ability to assess generated images across diverse criteria.

2.2 Search Algorithms

Formally, a search algorithm is some function $f : (\mathcal{V}, D_\theta, \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}) \mapsto \mathbf{z}^*$ which takes as input a verifier $\mathcal{V}$, a generative diffusion model D_θ, and N candidate noise latents. It then outputs the "best" noise latent $\mathbf{z}^*$ corresponding to the highest scoring image among N images sampled from D_θ, as determined by $\mathcal{V}$. Invoking f requires numerous forward passes through D_θ which constitutes the primary computational bottleneck; thus, this cost is quantified as the number of function evaluations (NFEs). Following Ma et al. [10], we consider two primary search strategies below.

Random Search. In random search, N candidate noise latents are sampled from a Gaussian distribution and progressively denoised over a fixed number of steps. A given verifier evaluates the resulting images, and the highest-scoring image among the N initial latents is returned. The scaling axis in random search is simply N; that is, NFEs $= N \cdot (\#\text{ of denoising steps})$. Notably, random search can lead to *verifier hacking*, a failure mode in which the search overfits to biases of the pretrained verifier rather than producing genuinely higher-quality generations [10].

Zero-Order Search. To mitigate verifier hacking, zero-order search traverses the noise space incrementally in directions that are locally optimal with respect to the verifier's score. Starting from an initial noise latent as the pivot, the algorithm samples $N - 1$ additional latents in its neighborhood. All N latents are denoised, and the latent yielding the highest-scoring image becomes the new pivot. This process is repeated for k iterations, resulting in NFEs $= kN \cdot (\#\text{ of denoising steps})$.

3 Method

In this work, we focus on the text-to-image generation task in the infrared domain, where, given text prompts, we aim to generate physically plausible infrared images with a pretrained text-to-image diffusion model. To that end, we apply inference-time scaling to guide the generation process toward higher-quality outputs that better reflect infrared imaging characteristics.

Infrared Image Generation. We select FLUX.1-dev [20], a state-of-the-art text-to-image model, as our diffusion model backbone. To endow the model with knowledge of the infrared domain, we first apply Low-Rank Adaptation (LoRA) [21] finetuning on a limited set of 1,000 infrared image–caption pairs. We observe that this lightweight adaptation is typically sufficient to equip the model with a basic understanding of infrared image characteristics; however, we find that the resulting generations exhibit high variability in quality, with many samples appearing physically implausible or resembling simple grayscale versions of RGB images (Figure 1). This inconsistent nature, which we attribute to inductive biases inherited from pretraining on large-scale RGB datasets, creates an opportunity to apply verifier-based inference-time selection to identify and promote higher-quality outputs.

Verifier Training. To steer generations towards more physically accurate infrared images, we adapt CLIP [13] as a zero-shot verifier. Since CLIP is pretrained on large-scale RGB image data, it is not suitable for evaluating the quality of infrared images. Accordingly, we finetune CLIP on paired infrared images and text captions, enabling the model to capture the unique characteristics of the infrared domain.

Motivated by our earlier observation that our diffusion model frequently generates grayscale-like images, we augment each training mini-batch with grayscale versions of the corresponding RGB

Table 1: Performance of two inference-time scaling methods on KAIST dataset.

Method	NFEs	α	Split	IRScore ($\times 10$) $\uparrow$	FID $\downarrow$
Naive Sampling	28	0.5	Test	0.1187	74.58
Random Search	336	0.5	Test	**0.6010**	**66.74**
Zero-Order Search	336	0.5	Test	0.4214	69.15

images. Further, we format the text captions as "`An INFRARED photo of {caption}.`" and "`A GRAYSCALE photo of {caption}.`" as appropriate. In doing so, we encourage the model to learn distinct representations for true infrared images and their grayscale counterparts.

Inference-Time Verification. Let $\Phi_I : \mathbb{R}^{h\times w} \to \mathbb{R}^d$ and $\Phi_T : V^{\leq n} \to \mathbb{R}^d$ denote the finetuned image and text encoders of the CLIP verifier, respectively. At inference time, given a candidate image $\mathbf{x} \in \mathbb{R}^{h\times w}$ generated by our finetuned diffusion model, along with its corresponding infrared and grayscale captions c_{IR} and c_{gray}, we compute an infrared quality score as

$$\text{IRS}\textsc{core}(\mathbf{x}) = (1-\alpha)\cos(\Phi_T(c_{\text{IR}}), \Phi_I(\mathbf{x})) - \alpha\cos(\Phi_T(c_{\text{gray}}), \Phi_I(\mathbf{x})), \tag{1}$$

where $\alpha \in [0, 1]$ is a tunable hyperparameter that balances alignment with true infrared semantics against undesired similarity to grayscale. This verifier-driven score serves as a selection criterion during sampling; we evaluate multiple candidate generations and retain those that maximize the score using either a random or zero-order search strategy, as described in Section 2.2.

4 Experiments

Setup. We conduct experiments on 49,561 images[1] from the KAIST Multispectral Pedestrian Detection Benchmark [22], a dataset of long-wave infrared and RGB pedestrian scenes captured from a vehicle. The data is divided into an 80/20 train–test split. As our diffusion backbone, we adopt FLUX.1-dev, a pretrained text-to-image model, augmented with a LoRA adapter of rank $r = 16$ finetuned on 1,000 infrared images from the KAIST training set along with corresponding captions. For verification, we employ CLIP-B/32 [13], finetuned on the same KAIST training data following the strategy outlined in Section 3.

Results. Our experiments suggest that inference-time scaling can result in meaningful gains in infrared generative quality. In particular, we observe a reduction in the Fréchet Inception Distance (FID) [23] score from 74.58 to 66.74 on the KAIST dataset (Table 1). The FID quantifies the discrepancy between the distribution of generated images and that of real images, with lower values indicating closer alignment; thus, this improvement provides promising evidence that inference-time scaling can enhance generative performance. Between the two search algorithms explored, random search achieves the largest improvement, reducing FID by over 10% relative to the naive baseline, while zero-order search achieves a more modest $\sim$7% gain under the same NFE budget. This aligns with the findings of Ma et al. [10] that the incremental update procedure of zero-order search results in slower convergence, as opposed to random search, which enables a broader exploration of the noise space. We also note that our NFE budget of 336 is relatively limited compared to other works that have explored inference-time scaling [10, 11], and we anticipate that performance could improve substantially by further scaling up the computational budget.

Further, we find qualitatively that our domain-adapted CLIP is effective in distinguishing between high and low-quality synthetic infrared images. As illustrated in Figure 1, the verifier assigns substantially higher scores to realistic infrared images that exhibit proper thermal characteristics, whereas an unmodified CLIP model fails to properly distinguish between such quality differences, instead assigning a higher score to the lower-quality image. We also report the mean of the IRScore (Equation 1) across all generated samples in Table 1, and find that both random and zero-order search effectively explore the noise space to identify latents with higher verifier scores, yielding consistent improvements over naive sampling.

[1] `https://huggingface.co/datasets/koifisharriet/KAIST-Multispectral-Pedestrian-Benchmark`

5 Conclusion

In this work, we introduced an inference-time scaling approach using a domain-adapted CLIP verifier for enhanced infrared image generation quality in text-to-image diffusion models. Our key contribution is training CLIP to distinguish between realistic infrared images and grayscale-like artifacts, enabling effective guidance of diffusion model sampling. Preliminary results on the KAIST dataset demonstrate FID score improvements of over 10% compared to unguided baselines, establishing inference-time guidance as a promising direction for infrared image generation in low-data settings. Building on this foundation, promising avenues for future research include training alternative verifiers such as physics-based or domain-specific models, as well as extending evaluation to other datasets and sensing modalities to assess the generality of the approach.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, December 2020. 1

[2] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations, February 2021. 1

[3] Lingyan Ran, Lidong Wang, Guangcong Wang, Peng Wang, and Yanning Zhang. DiffV2IR: Visible-to-Infrared Diffusion Model via Vision-Language Understanding, March 2025. 1

[4] Colin N. Reinhardt, Connor Anderson, and Elim Schenck. V2IR-CnLDM: A generative visible-to-infrared image translation using ControlNet-guided conditional latent diffusion model. *Optical Engineering*, 64(9):092206, June 2025. ISSN 0091-3286, 1560-2303. doi: 10.1117/1.OE.64.9.092206. 1

[5] Fangyuan Mao, Jilin Mei, Shun Lu, Fuyang Liu, Liang Chen, Fangzhou Zhao, and Yu Hu. PID: Physics-Informed Diffusion Model for Infrared Image Generation, June 2025. 2

[6] Jared Kaplan, Sam McCandlish, Tom Henighan, and Tom B Brown. Scaling Laws for Neural Language Models. 2

[7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January 2023. 2

[8] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models, December 2023. 2

[9] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters, August 2024. 2

[10] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps, January 2025. 2, 3, 4

[11] Vignav Ramesh and Morteza Mardani. Test-Time Scaling of Diffusion Models via Noise Trajectory Search, May 2025. 2, 4

[12] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A General Framework for Inference-time Scaling and Steering of Diffusion Models, January 2025. 2

[13] 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, February 2021. 2, 3, 4

[14] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems*, October 2022. 2

[15] Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Jaewon Min, Minjae Kim, Wooseok Jang, Hyoungwon Cho, Sayak Paul, SeonHwa Kim, Eunju Cha, Kyong Hwan Jin, and Seungryong Kim. A Noise is Worth Diffusion Guidance, December 2024. 2

[16] Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not All Noises Are Created Equally:Diffusion Noise Selection and Optimization, July 2024. 2

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. 3

[18] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, December 2022. 3

[19] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *Advances in Neural Information Processing Systems*, volume 36, pages 15903–15935, December 2023. 3

[20] Black-forest-labs/flux. black-forest-labs, August 2025. 3

[21] 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, October 2021. 3

[22] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, Boston, MA, USA, June 2015. IEEE. ISBN 978-1-4673-6964-0. doi: 10.1109/CVPR.2015.7298706. 4

[23] 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. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 4