FREQUENCY-ALIGNED LATENT ADAPTATION FOR TEXT-TO-INFRARED IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-infrared generation provides a controllable way to synthesize infrared images from text, offering a potential approach for expanding diverse infrared data resources. However, adapting pretrained text-to-image diffusion models to infrared imagery remains challenging because latent representations learned from natural images are not tailored to infrared-specific thermal structures and frequency characteristics. Directly modifying the latent space may improve infrared specialization but can disrupt the latent distribution expected by the denoising model, creating a trade-off between adaptation and latent compatibility. To address this issue, we propose Frequency-Aligned Latent Adaptation (FALA), a two-stage framework that learns infrared-aware latent corrections through frequency-aligned encoder adaptation and mean-only posterior refinement, then progressively transfers them to the denoising U-Net for stable adaptation. We further curate a large-scale infrared corpus from 69 datasets for text-to-infrared generation. Extensive experiments demonstrate that FALA consistently improves distribution fidelity and coverage across diverse infrared scenarios while preserving the pretrained diffusion prior without additional inference-time overhead.

1 INTRODUCTION

Infrared (IR) imagery provides complementary information beyond visible images and plays an important role in nighttime perception, surveillance, autonomous systems, and multispectral sensing (Zhao et al., 2023b;a; Liu et al., 2022). However, developing robust infrared perception models requires large-scale and diverse data, which remain difficult to obtain. Existing infrared datasets are typically collected for specific tasks, leading to limited scene diversity and inconsistent annotations. Moreover, video-based infrared datasets often contain substantial temporal redundancy. Therefore, generating diverse and semantically controllable infrared data remains an important challenge.

Recent text-to-image generative models provide a promising direction for addressing this problem. Large-scale generative foundation models achieve high-fidelity and semantic image synthesis (Xie et al., 2025b; Zhou et al., 2025; Fan et al., 2025), while autoregressive and flow-based approaches improve scalability and efficiency (Han et al., 2025; Pang et al., 2025; Wang et al., 2025b; Ren et al., 2025a;b). Their rich visual priors make them attractive for domain-specific generation. We study text-to-infrared generation, where infrared images are synthesized from text alone. Unlike visible-to-infrared translation requiring source images (Xiao et al., 2025; Lee et al., 2026), text-to-infrared generation enables flexible synthesis of infrared scenes specified by language.

However, adapting pretrained text-to-image models to infrared generation remains challenging. Existing diffusion adaptation strategies mainly rely on denoising-network fine-tuning, conditioning modules, or parameter-efficient updates (Ye et al., 2023; Li et al., 2024; Chen et al., 2024c; Mou et al., 2024). While effective for adapting generation behavior, these methods typically retain the pretrained latent representation. This limitation becomes critical for infrared generation, as the latent space learned from natural images may not sufficiently capture infrared-specific thermal structures and frequency characteristics, while latent modification can introduce distribution shifts that hinder the pretrained denoising U-Net.

To address these challenges, we propose Frequency-Aligned Latent Adaptation (FALA), a two-stage framework for text-to-infrared generation. In Stage I, the Infrared Frequency-Aligned Encoder (IR-FAE) learns frequency-aware latent corrections, while Latent-Compatible Posterior Learning

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

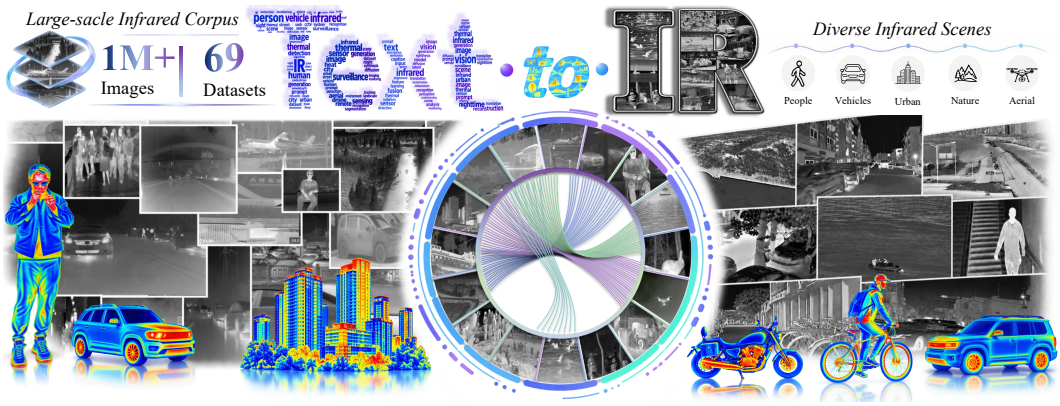


Figure 1: Overview of our large-scale infrared corpus and representative text-to-infrared generated samples. The corpus comprises approximately one million infrared images from 69 datasets, spanning diverse scenes, domains, and imaging conditions.

refines the posterior mean and preserves the pretrained posterior variance. In Stage II, Progressive Latent Adaptation (PLA) gradually introduces the learned latent correction into the denoising U-Net, enabling smooth adaptation toward the infrared-aligned latent distribution. We further construct a large-scale infrared corpus from 69 datasets and demonstrate consistent improvements in generation fidelity and coverage across diverse infrared scenarios.

Our contributions are summarized as follows:

- We identify two key challenges in text-to-infrared generation: infrared-specific latent representation and latent compatibility during diffusion adaptation.
- We propose FALA, which combines frequency-aware adaptation, posterior refinement, and progressive latent transition to adapt pretrained diffusion models for infrared generation.
- We construct a large-scale infrared corpus with diverse scenes and domains, and conduct comprehensive evaluations across diverse infrared scenarios.

2 Related Work

Text-to-Infrared Image Generation. Recent text-to-image foundation models, including LDM (Rombach et al., 2022), DiT (Peebles & Xie, 2023), SDXL (Podell et al., 2024), PixArt-α (Chen et al., 2024b), and SANA (Xie et al., 2025a), have demonstrated semantic generation capabilities for visible-spectrum images. However, these models are primarily trained on natural-image distributions and do not explicitly capture infrared-specific thermal characteristics. Existing infrared generation studies mainly focus on conditional generation settings. ThermalGen (Xiao et al., 2025) and TherA (Lee et al., 2026) perform visible-to-infrared translation, where a source visible image is required to preserve scene content. Other studies investigate infrared reconstruction (Ma et al., 2026) and super-resolution (Li et al., 2025) based on existing infrared observations. LanguageBind (Zhu et al., 2024a) aligns infrared data with natural language for cross-modal representation learning, but does not address image generation. Therefore, text-only infrared image generation, where language serves as the sole condition for synthesizing infrared scenes, remains relatively unexplored.

Domain Adaptation of Pretrained Diffusion Models. Pretrained text-to-image diffusion models are commonly adapted to new domains by transferring their semantic and generative priors through domain-specific fine-tuning or parameter-efficient updates (Duan et al., 2024; Zhu et al., 2024b). Existing approaches improve adaptation through denoising-network optimization, domain-aware guidance, or lightweight parameter updates. TGDP (Ouyang et al., 2024) introduces classifier guidance for source-to-target knowledge transfer under limited target-domain data, while Domain Guidance (Zhong et al., 2025) improves domain alignment through target-domain guidance during sampling. SaRA (Hu et al., 2025) and DEFT (Kumar et al., 2025) further explore efficient

parameter adaptation strategies. Despite their effectiveness, these approaches mainly focus on modifying the denoising process and typically assume that the pretrained latent representation remains transferable across domains. For substantial modality shifts such as visible-to-infrared adaptation, representation-level discrepancies may remain unresolved (Zhang et al., 2025), motivating latent-aware adaptation beyond conventional diffusion fine-tuning.

Latent and Frequency-Aware Representation. Latent diffusion models rely heavily on the quality of latent representations produced by image autoencoders. VAE (Kingma & Welling, 2014), VQ-VAE (van den Oord et al., 2017), and VQGAN (Esser et al., 2021) establish different strategies for constructing continuous or discrete generative latent spaces. Recent studies further investigate latent representations from the perspectives of compression efficiency, reconstruction quality, and semantic structure. DC-AE (Chen et al., 2025a) improves spatial compression efficiency for high-resolution generation, while VA-VAE (Yao et al., 2025) and LDMAE (Lee et al., 2025) explore improved trade-offs between reconstruction and generation. Beyond autoencoder design, REPA (Yu et al., 2025) aligns intermediate diffusion features with visual foundation representations, and RAE (Zheng et al., 2026) constructs semantically richer latent spaces using pretrained visual encoders. Meanwhile, frequency-aware generation methods such as Wavelet Diffusion (Phung et al., 2023) demonstrate the importance of frequency structures for image synthesis. However, existing latent-space studies mainly focus on improving general generation quality or redesigning latent representations, while adapting pretrained latent representations to infrared-specific thermal and frequency characteristics while preserving diffusion compatibility remains insufficiently explored.

3 METHOD

3.1 OVERVIEW

Adapting pretrained latent diffusion models to infrared (IR) generation requires improving infrared-specific latent representations while preserving compatibility with the pretrained denoising model. Existing adaptation methods mainly modify the denoising process, but overlook the representation gap between natural-image latent spaces and infrared thermal structures.

To address this issue, we propose **Frequency-Aligned Latent Adaptation (FALA)**, a two-stage framework for text-to-infrared generation. As shown in Fig. 2, Stage I learns an infrared-aware latent representation through the Infrared Frequency-Aligned Encoder (IR-FAE) and Latent-Compatible Posterior Learning. Specifically, IR-FAE introduces multi-scale frequency-aware residual adaptation into the frozen AutoencoderKL encoder to capture infrared-specific thermal structures, while latent mean refinement adjusts the posterior mean and retains the pretrained variance to limit changes to posterior uncertainty. Stage II performs Progressive Latent Adaptation (PLA), which gradually transitions the training latent from the base posterior toward the infrared-adapted representation, allowing the denoising U-Net to smoothly accommodate the latent distribution shift.

Given an infrared image-text pair (x, c), the frozen AutoencoderKL first produces the base posterior mean and log-variance $(\mu_{\text{base}}, \ell_{\text{base}})$. The FSA-augmented encoder and latent mean refiner then produce the infrared-adapted mean μ_{IR}, while retaining the base posterior variance. During U-Net adaptation, PLA progressively transitions from the base posterior mean toward the infrared-adapted mean. Only the newly introduced adaptation modules are optimized in Stage I, while only the denoising U-Net is optimized in Stage II.

3.2 INFRARED FREQUENCY-ALIGNED ENCODER

The pretrained AutoencoderKL encoder is optimized on natural-image distributions and does not explicitly model infrared-specific frequency characteristics. Infrared imagery contains complementary structures across frequency ranges: low-frequency components mainly describe spatially coherent thermal distributions, whereas high-frequency components preserve local variations and structural details. We therefore introduce the Infrared Frequency-Aligned Encoder (IR-FAE), which inserts Frequency-Split Adapters (FSAs) at three selected encoder scales. The pretrained encoder backbone remains frozen, while each FSA learns an infrared-specific residual correction to the corresponding intermediate feature.

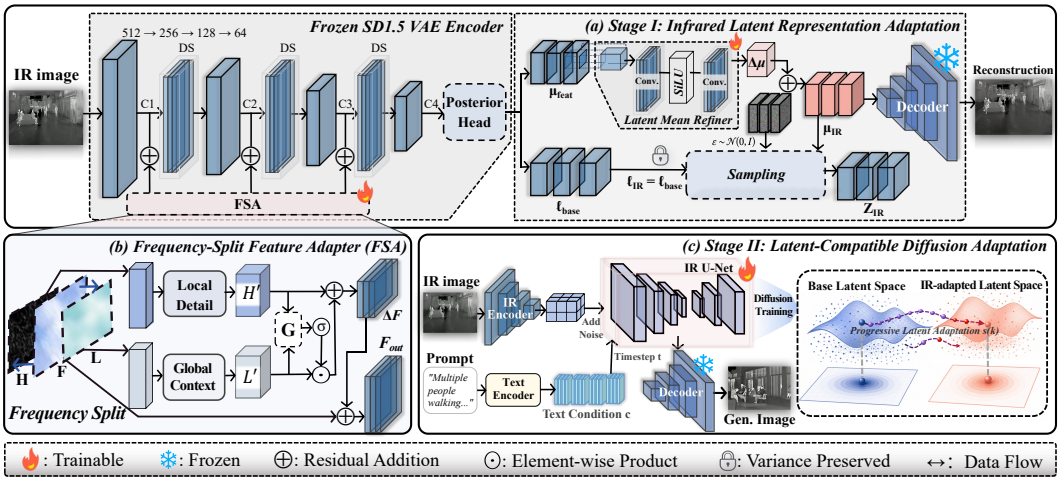


Figure 2: Overview of FALA. Stage I learns an infrared-aware latent representation through frequency-aligned encoder adaptation and latent-compatible mean refinement. Stage II progressively transitions the training latent from the base posterior toward the infrared-adapted representation to smoothly adapt the denoising U-Net.

Given an intermediate feature $h_l \in \mathbb{R}^{C_l \times H_l \times W_l}$, FSA forms a frequency-inspired decomposition using $h_l^{\text{low}} = \text{AvgPool}_{5\times5}(h_l)$ and $h_l^{\text{high}} = h_l - h_l^{\text{low}}$. This decomposition serves as an inductive bias rather than a strict physical frequency transform, allowing the two branches to separately model global thermal structure and local details.

Low-frequency global modeling. The low-frequency branch captures spatially coherent thermal structures beyond local receptive fields. We first project and resize h_l^{low} to $p_l^{\text{low}} = \text{Pool}_{16\times16}(\psi_l^{\text{low}}(h_l^{\text{low}}))$. Horizontal and vertical bidirectional GRUs aggregate directional dependencies, denoted by $s_l = \mathcal{G}_l(p_l^{\text{low}})$, followed by a lightweight bidirectional scan:

$$f_l^{\text{low}} = \mathcal{U}_l\left(\text{LN}\left[s_l + \text{BiScan}(s_l)\right] + p_l^{\text{low}}\right), \tag{1}$$

where $\mathcal{U}_l$ denotes upsampling and output projection to the original feature resolution.

Reliability-aware high-frequency modeling. The high-frequency branch preserves informative boundaries and local structures while suppressing unreliable responses. A lightweight local operator produces $\bar{f}_l^{\text{high}} = \mathcal{H}_l(h_l^{\text{high}})$. In parallel, we compute a spatial RMS energy map from the original high-frequency residual:

$$e_l^{\text{high}}(i,j) = \sqrt{\frac{1}{C_l}\sum_{c=1}^{C_l}\left(h_{l,c}^{\text{high}}(i,j)\right)^2 + \epsilon}, \qquad r_l = \sigma\Big(\mathcal{R}_l\left[\bar{f}_l^{\text{high}} \| e_l^{\text{high}}\right]\Big). \tag{2}$$

Here, $\|$ denotes channel concatenation and r_l is a learned single-channel reliability map. The refined high-frequency representation is $f_l^{\text{high}} = r_l \odot \bar{f}_l^{\text{high}}$.

Adaptive frequency fusion. The relative importance of global thermal structures and local details varies spatially. In addition to e_l^{high}, we compute the low-frequency response magnitude as $e_l^{\text{low}} = \text{Mean}_c(|h_l^{\text{low}}|)$. The fusion gate is conditioned on both branch features and their response statistics:

$$g_l = \sigma\Big(\mathcal{A}_l\left[f_l^{\text{low}} \| f_l^{\text{high}} \| e_l^{\text{low}} \| e_l^{\text{high}}\right]\Big). \tag{3}$$

The infrared-specific residual correction is

$$\Delta h_l = W_l^{\text{out}}\left[g_l \odot f_l^{\text{low}} + (1-g_l)\odot f_l^{\text{high}}\right], \qquad \widetilde{h}_l = h_l + \Delta h_l. \tag{4}$$

Applying FSAs across the selected encoder scales adapts the feature pathway while preserving the pretrained backbone.

3.3 LATENT-COMPATIBLE POSTERIOR LEARNING

Although IR-FAE improves infrared representation, directly using the fully adapted posterior may introduce a distribution shift to the pretrained denoising U-Net. We therefore refine the posterior mean while retaining the pretrained variance, enabling infrared-specific adaptation without simultaneously altering latent location and uncertainty.

The FSA-augmented encoder produces an intermediate latent mean μ_{feat} through the frozen posterior head. A lightweight Latent Mean Refiner R_ϕ operates only on the final four-channel posterior mean and predicts $\Delta\mu_\phi = R_\phi(\mu_{\text{feat}})$, yielding $\mu_{\text{IR}} = \mu_{\text{feat}} + \Delta\mu_\phi$. This complements the multi-scale feature adaptation of IR-FAE without modifying the frozen variance branch. Meanwhile, the frozen base encoder provides $(\mu_{\text{base}}, \ell_{\text{base}})$, where ℓ_{base} denotes the log-variance and $\sigma_{\text{base}} = \exp(0.5\ell_{\text{base}})$. The adapted posterior is defined as

$$q_{\text{IR}}(z|x) = \mathcal{N}\left(\mu_{\text{IR}}, \operatorname{diag}\left(\exp(\ell_{\text{base}})\right)\right). \tag{5}$$

Unlike jointly relearning both posterior moments, this asymmetric adaptation restricts the representation change to the latent location while preserving the uncertainty structure inherited from the pretrained VAE. It therefore allows infrared-specific latent movement without unnecessarily reshaping the stochastic interface expected by the denoising model.

During Stage I, only the FSAs and Latent Mean Refiner are optimized, while the AutoencoderKL backbone, posterior head, and decoder remain frozen. For deterministic Stage-I reconstruction, we decode the posterior modes as $\widehat{x}_{\text{IR}} = D_0(\mu_{\text{IR}})$ and $\widehat{x}_{\text{base}} = D_0(\mu_{\text{base}})$, respectively. The training objective is

$$\mathcal{L}_{\text{Stage-I}} = \lambda_{\text{img}}\mathcal{L}_{\text{img}} + \lambda_{\text{freq}}\mathcal{L}_{\text{freq}} + \lambda_{\text{lat}}\mathcal{L}_{\text{lat}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}}. \tag{6}$$

Here, the four terms respectively enforce reconstruction fidelity, frequency consistency, latent regularization, and balanced frequency fusion; their detailed formulations are described in the Appendix. The resulting infrared-adapted mean μ_{IR} serves as the target representation for progressive adaptation in Stage II.

3.4 PROGRESSIVE LATENT ADAPTATION

After Stage I, the infrared-adapted representation captures richer domain-specific structures but also shifts the latent distribution seen by the pretrained denoising U-Net. Directly exposing the U-Net to the fully adapted latent space from the beginning may therefore require it to simultaneously learn infrared appearance and compensate for an abrupt representation change. To alleviate this mismatch, we introduce Progressive Latent Adaptation (PLA), which gradually shifts the training latent from the pretrained representation toward the infrared-adapted one.

At optimization step k, PLA linearly increases the adaptation strength and interpolates between the base and infrared-adapted posterior means:

$$\rho_k = \rho_{\max} \min\left(\frac{k}{K_{\text{ramp}}}, 1\right), \qquad \mu_k = (1-\rho_k)\mu_{\text{base}} + \rho_k\mu_{\text{IR}}, \tag{7}$$

where K_{ramp} controls the transition duration and $\rho_{\max}$ determines the maximum adaptation strength. Here, k denotes the training optimization step rather than the diffusion timestep. Early training therefore remains close to the pretrained latent distribution, while later optimization progressively incorporates more infrared-specific information. Importantly, $\rho_{\max}$ does not require a complete transition to μ_{IR}; instead, it controls how strongly the infrared-adapted representation is introduced during U-Net training.

Since Stage I retains the base posterior variance, the clean latent used for diffusion training is sampled from the progressively adapted posterior as

$$z_0^{(k)} = \gamma\left(\mu_k + \sigma_{\text{base}} \odot \xi\right), \qquad \xi \sim \mathcal{N}(0, I), \tag{8}$$

where γ is the fixed latent scaling factor of Stable Diffusion v1.5 and ξ denotes the VAE posterior sampling noise. The sampled clean latent is then perturbed by the standard diffusion forward process, $z_t^{(k)} = \sqrt{\bar{\alpha}_t} z_0^{(k)} + \sqrt{1-\bar{\alpha}_t}\epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$, and paired with the text embedding $T(c)$ from the frozen text encoder for U-Net training. PLA therefore modifies only the clean-latent transition path while leaving the diffusion objective and noise formulation unchanged.

Table 1: Quantitative comparison of text-to-infrared adaptation strategies. (a) Overall performance on Test10K. (b) Scene-level performance across five infrared scenarios. KID is reported as $\times 10^3$, and DINO denotes FD-DINOv2. Light pink and light blue indicate the best and second-best results, respectively.

(a) Overall performance on Test10K

Method	Source	FID↓	KID↓	CMMD↓	DINO↓	Coverage↑
SD1.5+LoRA (Rombach et al., 2022)	CVPR'22	56.18	12.41	37.41	522.64	0.1114
SDXL+LoRA (Podell et al., 2024)	ICLR'24	59.51	24.34	64.13	561.03	0.1061
PixArt-Σ+LoRA (Chen et al., 2024a)	ECCV'24	137.30	105.45	98.11	837.06	0.0525
FLUX+LoRA	BFL'24	93.76	46.49	188.93	1133.96	0.0257
Thermal LoRA (dvyio, 2024)	HF Repo	83.72	40.46	135.21	1004.83	0.0315
SD1.5 FT (Rombach et al., 2022)	CVPR'22	60.23	13.99	39.04	535.78	0.1158
SDXL FT (Podell et al., 2024)	ICLR'24	90.35	47.09	173.39	979.98	0.0335
Ours	**Proposed**	**53.41**	**10.42**	**37.22**	**511.41**	**0.1162**

(b) Scene-level performance

Method	Metrics	Ground	Aerial	Surveillance	Small Target	Multispec.
SD1.5-L	FID/KID	82.02/20.71	96.59/30.82	79.37/19.93	78.72/25.57	74.86/21.64
	CMMD/DINO	45.31/772.16	64.14/770.71	51.93/662.50	63.79/736.29	49.12/787.12
SD1.5-FT	FID/KID	83.61/15.90	**89.30/21.05**	72.62/12.91	79.22/21.28	76.71/17.72
	CMMD/DINO	52.53/811.03	50.57/734.12	49.18/603.00	61.54/643.24	51.24/782.23
SDXL-L	FID/KID	92.28/35.76	100.06/42.17	80.41/28.69	75.88/31.81	80.50/32.78
	CMMD/DINO	109.78/887.48	77.49/774.46	70.96/623.57	87.20/716.64	97.86/846.05
T-LoRA	FID/KID	103.70/49.20	122.12/55.55	105.88/43.29	97.65/50.45	87.23/38.29
	CMMD/DINO	173.73/1345.69	159.08/1243.86	132.53/1071.72	189.37/1155.30	153.83/1233.72
Ours	**FID/KID**	**69.68/9.98**	93.42/21.92	**65.12/10.13**	**66.36/13.26**	**66.58/12.02**
	CMMD/DINO	**45.21/769.70**	**50.01/729.84**	**45.40/550.07**	**58.02/613.90**	**47.90/755.58**

During Stage II, IR-FAE, the AutoencoderKL, and the text encoder remain frozen, while only the denoising U-Net is optimized. This progressive curriculum allows the U-Net to adapt to the infrared-specific representation without an abrupt latent shift while retaining the semantic prior inherited from pretraining. At inference time, FALA follows the standard latent diffusion pipeline using the adapted U-Net together with the frozen text encoder and decoder, without additional inference-time modules or computational overhead.

4 Experiments

4.1 Experimental Settings

Implementation Details. We implement FALA based on Stable Diffusion v1.5 (Rombach et al., 2022) at 512×512 resolution. In Stage I, only the Frequency-Split Adapters in IR-FAE and the Latent Mean Refiner are optimized, while the VAE backbone, posterior head, and decoder remain frozen. In Stage II, we freeze the VAE, IR-FAE, Latent Mean Refiner, and text encoder, and optimize only the denoising U-Net. The latent adaptation strength ρ_k increases linearly from 0 to 0.5 over the first 1,250 optimization steps and remains fixed thereafter. All models are trained on NVIDIA GeForce RTX 5090 GPUs.

Datasets and Metrics. We curate a large-scale infrared corpus from 69 publicly available infrared datasets, containing 1,092,778 training images, 5,000 validation images, and 10,000 test images. Validation5K and Test10K are constructed using leakage-safe splits with no overlap across training, validation, and test partitions. For generation evaluation, we report FID (Heusel et al., 2017), KID (Bińkowski et al., 2018), CMMD (Jayasumana et al., 2024), FD-DINOv2 (Stein et al., 2023), and Coverage (Naeem et al., 2020) to evaluate distribution fidelity and coverage. For Stage-I reconstruction and encoder ablations, we additionally report PSNR, SSIM (Wang et al., 2004),

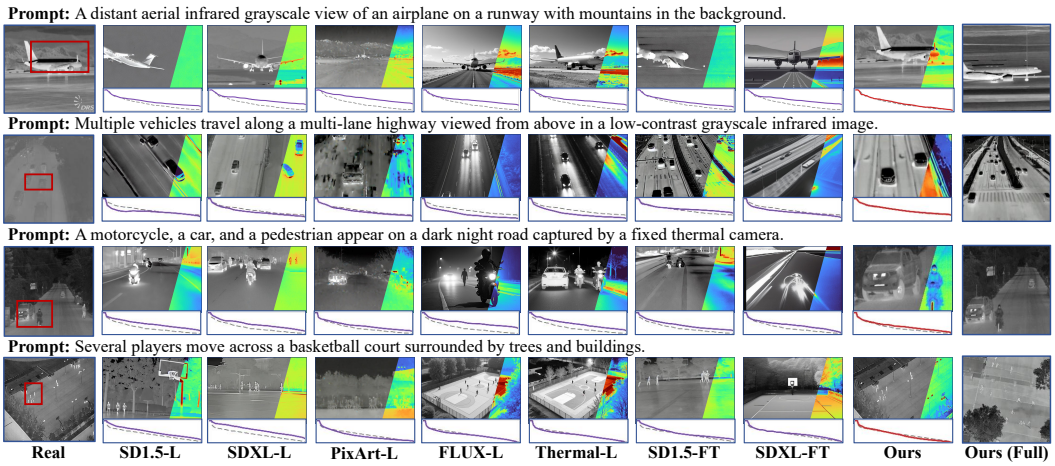


Figure 3: Qualitative comparison of text-to-infrared generation. The first column shows real infrared references and selected ROIs, followed by generated results from different methods. Curves below the ROIs show radially averaged power spectra, with real infrared images as reference.

LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), Spectral Error, and Band Energy Error to evaluate spatial reconstruction quality and frequency consistency.

4.2 Comparison with Existing Methods

Compared Methods. We compare FALA with representative parameter-efficient and full fine-tuning strategies for text-to-infrared generation. For parameter-efficient adaptation, LoRA (Hu et al., 2022) is applied to Stable Diffusion v1.5 (Rombach et al., 2022), SDXL (Podell et al., 2024), PixArt-Σ (Chen et al., 2024a), and FLUX.1-dev. We additionally evaluate full U-Net fine-tuning of SD1.5 and SDXL, together with a publicly released FLUX.1-dev thermal-image LoRA initialization (dvyio, 2024). All methods are adapted on the same infrared training corpus and evaluated on Test10K using the same prompts, output resolution, and evaluation protocol.

Quantitative Evaluation. As shown in Table 1(a), FALA achieves the best overall performance on Test10K across all evaluated metrics. Compared with SD1.5+LoRA, the strongest parameter-efficient baseline, FALA further reduces FID and KID while improving CMMD, FD-DINOv2, and Coverage. These consistent gains show that explicitly adapting the latent representation provides complementary benefits beyond denoiser-only adaptation, improving infrared distribution fidelity while maintaining distribution coverage.

Table 1(b) further evaluates the methods across five representative infrared scenarios. FALA achieves the best results across all reported metrics on ground, surveillance, small-target, and multispectral scenes. On aerial imagery, it obtains the best CMMD and FD-DINOv2 while remaining competitive in FID and KID. These results indicate that the improvements extend across heterogeneous infrared distributions rather than being concentrated in a particular scene category.

Qualitative Analysis. Figure 3 further compares the spatial and frequency characteristics of generated infrared images. Although existing methods generally preserve coarse scene semantics, they often weaken small thermal targets, blur object boundaries, or produce inconsistent local intensity structures. In contrast, FALA better preserves coherent thermal distributions and fine local structures. The radial power spectra of selected ROIs also more closely follow those of real infrared references, providing complementary evidence that the generated images better match real infrared characteristics in both spatial and frequency domains.

Compositional Text Following. Figure 4 evaluates fine-grained text following under challenging prompts involving human interactions, relative spatial relationships, and partial occlusion. While several baselines generate the requested subjects, they often fail to satisfy all relational constraints

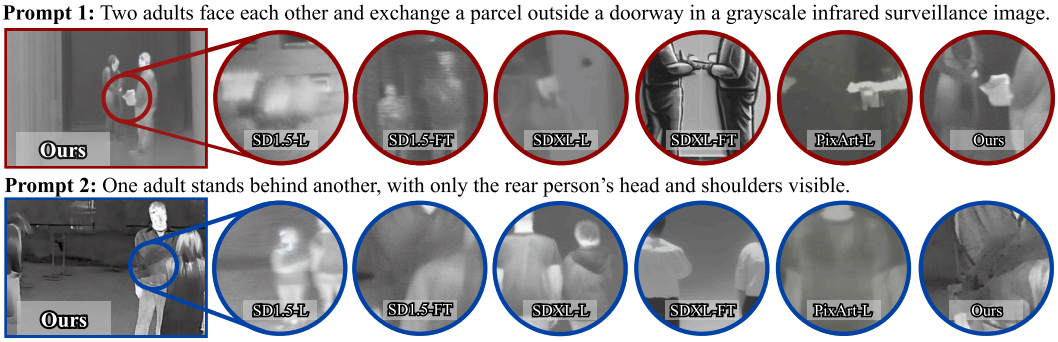


Figure 4: **Comparison on challenging compositional prompts.** Enlarged regions highlight the key relational constraints in each prompt.

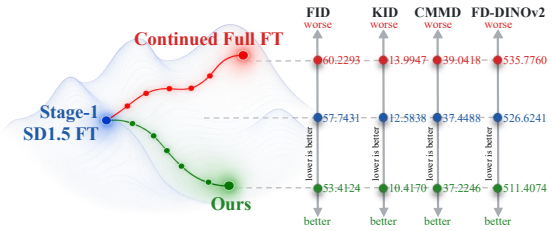


Figure 5: Stage-II adaptation trajectories.

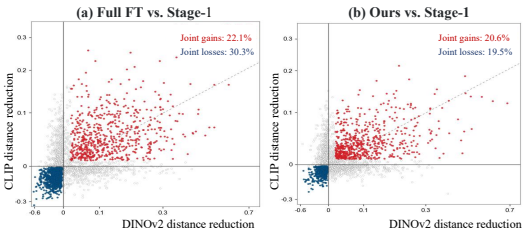


Figure 6: Natural-to-IR feature discrepancy.

simultaneously. FALA more consistently preserves the specified interactions and spatial configurations, suggesting that infrared-domain adaptation retains the compositional capability inherited from the pretrained generative prior.

4.3 Ablation Study

We analyze three key aspects of FALA: the necessity of infrared-aware latent adaptation, the frequency-aware latent representation learned in Stage I, and the progressive latent transition used in Stage II. These studies systematically validate the motivation of latent adaptation and the key design choices across both training stages of FALA.

Effectiveness of Infrared-Aware Latent Adaptation. We first examine whether the improvement can be obtained simply by continuing to optimize the denoising U-Net. As shown in Fig. 5, continued full U-Net fine-tuning gradually degrades generation performance over the evaluated training interval, whereas incorporating the infrared-adapted latent produces a consistent improvement. This result indicates that the remaining limitation cannot be addressed by denoiser optimization alone. The feature-space visualization in Fig. 6 further shows a reduced discrepancy between natural-image and infrared representations after IR-FAE adaptation. Together, these observations support the motivation of FALA that adapting the latent representation provides complementary benefits beyond conventional denoiser-side fine-tuning.

Stage-I Frequency-Aware Latent Learning. Figure 7 evaluates the main design choices of IR-FAE. Removing explicit frequency decomposition degrades both spatial reconstruction and frequency consistency, while alternative decompositions based on Gaussian filtering, Haar wavelets, FFT, and local filtering are less effective than the proposed design. These results indicate that directly learning mixed-frequency features is less suitable for modeling the heterogeneous thermal structures of infrared imagery. Moreover, low-frequency-only and high-frequency-only variants both underperform the complete model, confirming that global thermal distributions and local structural details provide complementary information. Adaptive fusion further improves over fixed fusion, showing that the relative contributions of the two frequency branches should vary with spatial content rather than being combined uniformly.

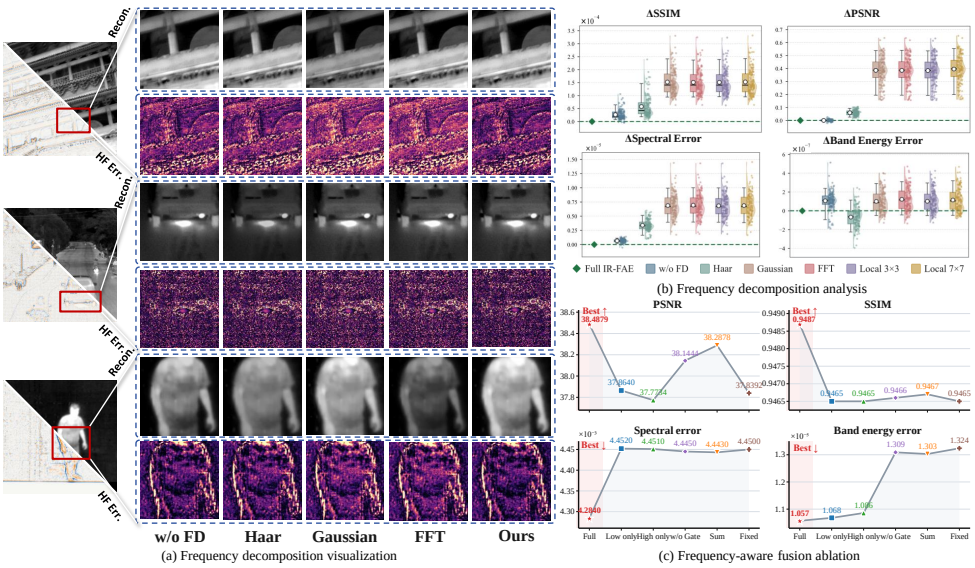


Figure 7: Ablation of frequency-aware latent learning in Stage I. (a) Frequency decomposition. (b) Spatial-frequency evaluation. (c) Frequency-branch and fusion ablations. Positive Δ denotes degradation.

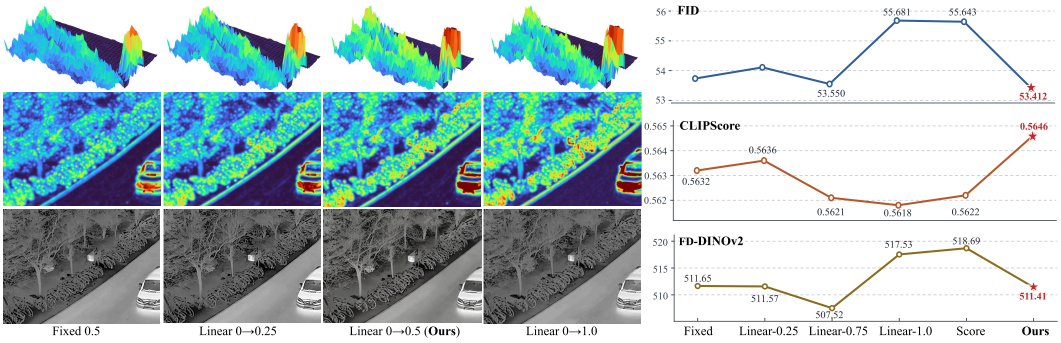


Figure 8: Ablation of progressive latent adaptation in Stage II. Effects of the latent transition strategy and maximum adaptation strength $\rho_{\max}$ on generation performance.

Stage-II Progressive Latent Adaptation. Figure 8 analyzes how the infrared-specific latent correction is introduced during U-Net adaptation. Directly exposing the denoiser to the adapted latent from the beginning yields inferior performance to the progressive transition, suggesting that an abrupt latent distribution shift hinders adaptation. We further vary the maximum adaptation strength $\rho_{\max}$ to examine the extent of latent correction. Among the evaluated settings, $\rho_{\max} = 0.5$ achieves the most favorable overall performance, whereas weaker or stronger corrections are less effective. This trend supports progressively introducing a controlled latent correction to balance infrared-specific representation learning with compatibility to the pretrained latent distribution.

5 Conclusion

We propose FALA, a two-stage framework for text-to-infrared generation that adapts latent representations while preserving compatibility with pretrained diffusion models. FALA combines frequency-aware infrared encoding, latent-compatible posterior refinement, and progressive latent adaptation to bridge the natural-to-infrared representation gap. Experiments demonstrate consistent gains across diverse infrared scenarios without inference-time overhead. Future work will explore stronger backbones and broader sensing conditions.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation to improve textual clarity, organization, and readability, as well as to assist with the refinement of scientific figures and overall manuscript presentation. All AI-assisted suggestions were carefully reviewed and manually revised by the authors based on the original experiments, data, and references. The authors take full responsibility for the accuracy, integrity, and conclusions of the submitted manuscript.

ETHICS STATEMENT

Our study relies on publicly available infrared datasets and does not involve new data collection from human subjects. All datasets are used solely for academic research purposes and are processed in accordance with their respective terms of use. The infrared images used in this work do not intentionally expose identifiable personal information. We take measures to minimize potential privacy risks throughout data preparation, model training, and evaluation.

REPRODUCIBILITY STATEMENT

The main paper and appendix provide detailed descriptions of the proposed framework, including the model architecture, training objectives, implementation details, dataset protocols, evaluation settings, and additional experimental results. These descriptions are provided to facilitate the reproduction and verification of our experiments.

REFERENCES

Chris H. Bahnsen and Thomas B. Moeslund. Rain removal in traffic surveillance: Does it matter? *IEEE Transactions on Intelligent Transportation Systems*, 20(8):2802–2819, 2019. doi: 10.1109/TITS.2018.2872502. Paper associated with the AAU RainSnow dataset.

Yannick Benezeth, Désiré Sidibé, and Jean-Baptiste Thomas. Background subtraction with multispectral video sequences. In *12th OMNIVIS Workshop of the IEEE International Conference on Robotics and Automation*, 2014. URL `https://hal.science/hal-00986168`. Original publication associated with the MSVS dataset.

Amanda Berg, Jörgen Ahlberg, and Michael Felsberg. A thermal object tracking benchmark. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 1–6, 2015. doi: 10.1109/AVSS.2015.7301772.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

Elizabeth Bondi, Raghav Jain, Palash Aggrawal, Saket Anand, Robert Hannaford, Ashish Kapoor, Jim Piavis, Shital Shah, Lucas Joppa, Bistra Dilkina, and Milind Tambe. BIRDSAI: A dataset for detection and tracking in aerial thermal infrared videos. In *2020 IEEE Winter Conference on Applications of Computer Vision*, pp. 1747–1756, 2020. doi: 10.1109/WACV45572.2020.9093284.

Georgios Botsoglou, Marios Lysitsas, Dimitris Dimitriadis, Constantina Tsokana, George Valiakos, and Grigorios Tsoumakas. TIDS: A thermal imaging dataset for subclinical mastitis in dairy sheep. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, Lecture Notes in Computer Science, pp. 408–425, 2025. doi: 10.1007/978-3-032-06118-8_24. Use this only if TIDS denotes the dairy-sheep thermal dataset.

Martin Brenner, Napoleon H. Reyes, Teo Susnjak, and Andre L. C. Barczak. MM5: Multimodal image capture and dataset generation for RGB, depth, thermal, UV, and NIR. *Information Fusion*, 126:103516, 2026. doi: 10.1016/j.inffus.2025.103516.

Matthew Brown and Sabine Süsstrunk. Multi-spectral SIFT for scene category recognition. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 177–184, 2011. doi: 10.1109/CVPR.2011.5995637.

Zhe Cao, Jin Zhang, and Ruiheng Zhang. IRGPT: Understanding real-world infrared image with bi-cross-modal curriculum on large-scale benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 166–176, 2025.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*, 2024a.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024b.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *ICLR*, 2025a.

Junzhang Chen and Xiangzhi Bai. Atmospheric transmission and thermal inertia induced blind road segmentation with a large-scale dataset TBRSD. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1053–1063, 2023. doi: 10.1109/ICCV51070.2023.00103.

Junzhang Chen, Zichao Liu, Darui Jin, Yuanyuan Wang, Fan Yang, and Xiangzhi Bai. Light transport induced domain adaptation for semantic segmentation in thermal infrared urban scenes. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):23194–23211, 2022. doi: 10.1109/TITS.2022.3194931.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. AnyDoor: Zero-shot object-level image customization. In *CVPR*, 2024c.

Zizhao Chen, Yeqiang Qian, Xiaoxiao Yang, Chunxiang Wang, and Ming Yang. AMFD: Distillation via adaptive multimodal fusion for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 27:8298–8310, 2025b. doi: 10.1109/TMM.2025.3604937. Paper introducing the SMOD dataset.

Yimian Dai, Xiang Li, Fei Zhou, Yulei Qian, Yaohong Chen, and Jian Yang. One-stage cascade refinement networks for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023. doi: 10.1109/TGRS.2023.3243062. Paper introducing SIRST-V2.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Yuxuan Duan, Yan Hong, Bo Zhang, Jun Lan, Huijia Zhu, Weiqiang Wang, Jianfu Zhang, Li Niu, and Liqing Zhang. DomainGallery: Few-shot domain-driven image generation by attribute-centric finetuning. In *NeurIPS*, 2024.

dvyio. Flux lora thermal image. Hugging Face, 2024. URL `https://huggingface.co/dvyio/flux-lora-thermal-image`. Accessed September 3, 2026.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. In *ICLR*, 2025.

FLIR Systems. FLIR thermal dataset for advanced driver assistance systems. Online dataset, 2018. URL `https://www.flir.com/oem/adas/adas-dataset-form/`. Accessed September 25, 2026.

James E. Gallagher and Edward J. Oughton. Assessing thermal imagery integration into object detection methods on air-based collection platforms. *Scientific Reports*, 13:8491, 2023. doi: 10.1038/s41598-023-34791-8. Associated RGB-LWIR air-based dataset: 10.5281/zenodo.7465521.

Alejandro González, Zhijie Fang, Yainuvis Socarrás, Joan Serrat, David Vázquez, Jiaolong Xu, and Antonio M. López. Pedestrian detection at day/night time with visible and FIR cameras: A comparison. *Sensors*, 16(6):820, 2016. doi: 10.3390/s16060820.

Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5108–5115, 2017. doi: 10.1109/IROS.2017.8206396.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, 2025.

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

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, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Teng Hu, Jiangning Zhang, Ran Yi, Hongrui Huang, Yabiao Wang, and Lizhuang Ma. SaRA: High-efficient diffusion model fine-tuning with progressive sparse low-rank adaptation. In *ICLR*, 2025.

Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1037–1045, 2015.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. LLVIP: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3496–3504, 2021.

Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Jian Zhao, Guodong Guo, and Zhenjun Han. Anti-UAV: A large multi-modal benchmark for UAV tracking, 2021. Citation for the RGB-T multimodal Anti-UAV benchmark.

Zhiying Jiang, Zengxi Zhang, Jinyuan Liu, Xin Fan, and Risheng Liu. Multispectral image stitching via global-aware quadrature pyramid regression. *IEEE Transactions on Image Processing*, 33: 4288–4302, 2024. doi: 10.1109/TIP.2024.3430532.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Komal Kumar, Rao Anwer, Fahad Shahbaz Khan, Salman Khan, Ivan Laptev, and Hisham Cholakkal. DEFT: Decompositional efficient fine-tuning for text-to-image models. In *NeurIPS*, 2025.

Connor Lee, Matthew Anderson, Nikhil Ranganathan, Xingxing Zuo, Kevin Do, Georgia Gkioxari, and Soon-Jo Chung. Caltech aerial RGB-thermal dataset in the wild. In *Computer Vision – ECCV 2024*, pp. 236–256, 2024. doi: 10.1007/978-3-031-73036-8_14.

Dong-Guw Lee, Tai Hyoung Rhee, Hyunsoo Jang, Young-Sik Shin, Ukcheol Shin, and Ayoung Kim. TherA: Thermal-aware visual-language prompting for controllable rgb-to-thermal infrared translation. In *CVPR*, 2026.

Junho Lee, Jeongwoo Shin, Hyungwook Choi, and Joonseok Lee. Latent diffusion models with masked autoencoders. In *ICCV*, 2025.

Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 1856–1864, 2017. doi: 10.1145/3123266.3123289.

Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):3069–3082, 2021. doi: 10.1109/TNNLS.2020.3009373.

Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. LasHeR: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2022. doi: 10.1109/TIP.2021.3130533.

Xingyuan Li, Zirui Wang, Yang Zou, Zhixin Chen, Jun Ma, Zhiying Jiang, Long Ma, and Jinyuan Liu. DifIISR: A diffusion model with gradient guidance for infrared image super-resolution. In *CVPR*, 2025.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. PhotoMaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024.

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*, pp. 19800–19808, 2022.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *CVPR*, 2022.

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

Yingjian Liu, Gangnian Zhao, Shuzhen Fan, Cheng Fei, Junliang Liu, Zhishuo Zhang, Liqian Wang, Yongfu Li, Xian Zhao, and Zhaojun Liu. Tri-band vehicle and vessel dataset for artificial intelligence research. *Scientific Data*, 12:592, 2025. doi: 10.1038/s41597-025-04945-6.

Tengyu Ma, Zhilong Dai, Yubo Diao, Guanming An, Long Ma, Jinyuan Liu, and Risheng Liu. Taming generative diffusion model for task-oriented infrared imaging. In *CVPR*, 2026.

Jonas Mirlach, Lei Wan, Andreas Wiedholz, Hannan Ejaz Keen, and Andreas Eich. R-LiViT: A LiDAR-visual-thermal dataset enabling vulnerable road user focused roadside perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28375–28384, 2025.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020.

Junlin Ouyang, Qingwang Wang, Ying Shang, Pengcheng Jin, Hangwei Zhong, Liman Zhou, and Tao Shen. An RGB-TIR dataset from UAV platform for robust urban traffic scenes semantic segmentation. *Scientific Data*, 12:1701, 2025. doi: 10.1038/s41597-025-05994-7.

Yidong Ouyang, Liyan Xie, Hongyuan Zha, and Guang Cheng. Transfer learning for diffusion models. In *NeurIPS*, 2024.

Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T. Freeman, and Yu-Xiong Wang. RandAR: Decoder-only autoregressive visual generation in random orders. In *CVPR*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Tao Peng, Qing Li, and Pengfei Zhu. RGB-T crowd counting from drone: A benchmark and MMCCN network. In *Computer Vision – ACCV 2020*, volume 12627 of *Lecture Notes in Computer Science*, pp. 497–513, 2021. doi: 10.1007/978-3-030-69544-6_30.

Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *CVPR*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

Sébastien Razakarivony and Frédéric Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016. doi: 10.1016/j.jvcir.2015.11.002. Dataset is conventionally styled VEDAI.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. FlowAR: Scale-wise autoregressive image generation meets flow matching. In *ICML*, 2025a.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-x prediction for autoregressive visual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15781–15791, 2025b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Julian Ruck, Elena Vollmer, Rebekka Volk, and Marinus Vogl. Thermal anomaly segmentation dataset – thermal UAS-based images from germany with annotations for semantic segmentation model training, 2025.

Shreyas S. Shivakumar, Neil Rodrigues, Alex Zhou, Ian D. Miller, Vijay Kumar, and Camillo J. Taylor. PST900: RGB-thermal calibration, dataset and segmentation network. In *2020 IEEE International Conference on Robotics and Automation*, pp. 9441–9447, 2020. doi: 10.1109/ICRA40945.2020.9196831.

Kechen Song, Hongwei Wen, Xiaotong Xue, Liming Huang, Yingying Ji, and Yunhui Yan. Modality registration and object search framework for UAV-based unregistered RGB-T image salient object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. doi: 10.1109/TGRS.2023.3332179. Paper introducing UAV RGB-T 2400.

Simon Speth, Artur Gonçalves, Bastien Rigault, Satoshi Suzuki, Mondher Bouazizi, Yutaka Matsuo, and Helmut Prendinger. Deep learning with RGB and thermal images onboard a drone for monitoring operations. *Journal of Field Robotics*, 39(6):840–868, 2022. doi: 10.1002/rob.22082.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Christian Stippel, Thomas Heitzinger, and Martin Kampel. A trimodal dataset: RGB, thermal, and depth for human segmentation and temporal action detection. In *Pattern Recognition*, pp. 18–33, 2024. doi: 10.1007/978-3-031-54605-1_2.

Jiashun Suo, Tianyi Wang, Xingzhou Zhang, Haiyang Chen, Wei Zhou, and Weisong Shi. HIT-UAV: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection. *Scientific Data*, 10:227, 2023. doi: 10.1038/s41597-023-02066-6.

Mengyu Tan, Wentao Chao, Jo-Ku Cheng, Mo Zhou, Yiwen Ma, Xinyi Jiang, Jianping Ge, Lian Yu, and Limin Feng. Animal detection and classification from camera trap images using different mainstream object detection architectures. *Animals*, 12(15):1976, 2022. doi: 10.3390/ani12151976.

Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42, 2022. doi: 10.1016/j.inffus.2021.12.004.

Alexander Toet. TNO image fusion dataset, 2014.

Alexander Toet, Maarten A. Hogervorst, and Alan R. Pinkus. The TRICLOBS dynamic multi-band image data set for the development and evaluation of image fusion methods. *PLOS ONE*, 11(12): e0165016, 2016. doi: 10.1371/journal.pone.0165016.

Thao-Anh Tran, Vu-Minh Le, Thanh-Tung Phan, Dung Hoang, Duc Phan, Huong Ninh, and Hai Tran. Advancing thermal multi-object tracking with attention and metric fusion. *Open Computer Science*, 16(1):20250055, 2026. doi: 10.1515/comp-2025-0055.

Zhengzheng Tu, Tian Xia, Chenglong Li, Xiaoxiao Wang, Yan Ma, and Jin Tang. RGB-T image saliency detection via collaborative graph learning. *IEEE Transactions on Multimedia*, 22(1): 160–173, 2020. doi: 10.1109/TMM.2019.2924578.

Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. RGBT salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 25: 4163–4176, 2023. doi: 10.1109/TMM.2022.3171688.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

Jue Wang, Yuxiang Lin, Qi Zhao, Dong Luo, Shuaibao Chen, Wei Chen, and Xiaojiang Peng. Invisible gas detection: An RGB-thermal cross attention network and a new benchmark. *Computer Vision and Image Understanding*, 248:104099, 2024. doi: 10.1016/j.cviu.2024.104099. Paper introducing Gas-DB.

Kunpeng Wang, Keke Chen, Chenglong Li, Zhengzheng Tu, and Bin Luo. Alignment-free RGB-T salient object detection: A large-scale dataset and progressive correlation network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):7780–7788, 2025a. doi: 10.1609/aaai.v39i7.32838. Dataset name is UVT20K, not UTV20K.

Yadong Wang. Turbulence-distorted infrared imaging dataset, 2023.

Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han, Haoyuan Guo, Zhenheng Yang, Difan Zou, Jiashi Feng, and Xihui Liu. Parallelized autoregressive visual generation. In *CVPR*, 2025b.

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

Zheng Wu, Nathan Fuller, Diane Theriault, and Margrit Betke. A thermal infrared video benchmark for visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 201–208, 2014. doi: 10.1109/CVPRW.2014.39.

Jiuhong Xiao, Roshan Nayak, Ning Zhang, Daniel Tortei, and Giuseppe Loianno. ThermalGen: Style-disentangled flow-based generative models for rgb-to-thermal image translation. In *NeurIPS*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution image synthesis with linear diffusion transformers. In *International Conference on Learning Representations*, 2025a.

Housheng Xie, Meng Sang, Yukuan Zhang, Yang Yang, Shan Zhao, and Jianbo Zhong. RCVS: A unified registration and fusion framework for video streams. *IEEE Transactions on Multimedia*, 26:11031–11043, 2024. doi: 10.1109/TMM.2024.3443673. Paper introducing the HDO dataset.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025b.

Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. FusionDN: A unified densely connected network for image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12484–12491, 2020. doi: 10.1609/aaai.v34i07.6936.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-IF: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27026–27035, 2024. Paper introducing the EMS benchmark; also cite the original source datasets for any reused subsets.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*, 2025.

M. M. Zhang, J. Choi, K. Daniilidis, M. T. Wolf, and C. Kanan. VAIS: A dataset for recognizing maritime imagery in the visible and infrared spectrums. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 10–16, 2015. doi: 10.1109/CVPRW.2015.7301291.

Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. ISNet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 877–886, 2022.

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

Tao Zhang, Jinyong Wen, Zhen Chen, Kun Ding, Shiming Xiang, and Chunhong Pan. UNIP: Rethinking pre-trained attention patterns for infrared semantic segmentation. In *ICLR*, 2025.

Yongle Zhang and Zhanchuan Cai. CE-RetinaNet: A channel enhancement method for infrared wildlife detection in UAV images. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–12, 2023. doi: 10.1109/TGRS.2023.3299651. Associated publication specified by the ISOD dataset record.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *CVPR*, 2023a.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. DDFM: Denoising diffusion model for multi-modality image fusion. In *ICCV*, 2023b.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *ICLR*, 2026.

Jincheng Zhong, Xiangcheng Zhang, Jianmin Wang, and Mingsheng Long. Domain guidance: A simple transfer approach for a pre-trained diffusion model. In *ICLR*, 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Zhang, Zhifeng Li, Wei Liu, and Yuan Li. LanguageBind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*, 2024a.

Jingyuan Zhu, Shiyu Li, Yuxuan Liu, Jian Yuan, Ping Huang, Jiulong Shan, and Huimin Ma. ODGEN: Domain-specific object detection data generation with diffusion models. In *NeurIPS*, 2024b.

Yabin Zhu, Qianwu Wang, Chenglong Li, Jin Tang, Chengjie Gu, and Zhixiang Huang. Visible–thermal multiple object tracking: Large-scale video dataset and progressive fusion approach. *Pattern Recognition*, 161:111330, 2025. doi: 10.1016/j.patcog.2024.111330.

Yang Zou, Jun Ma, Zhidong Jiao, Xingyuan Li, Zhiying Jiang, and Jinyuan Liu. Toward real-world infrared image super-resolution: A unified autoregressive framework and benchmark dataset. In *CVPR*, 2026a.

Yang Zou, Xingyue Zhu, Kaiqi Han, Jun Ma, Xingyuan Li, Zhiying Jiang, and Jinyuan Liu. HATIR: Heat-aware diffusion for turbulent infrared video super-resolution. In *AAAI*, 2026b.

A LARGE-SCALE INFRARED CORPUS CONSTRUCTION

A.1 CORPUS OVERVIEW

Existing infrared datasets are primarily developed for task-specific applications, such as detection, tracking, fusion, and recognition. As a result, they often contain limited scene diversity, heterogeneous imaging characteristics, and inconsistent data formats, which makes them challenging to directly apply for training large-scale text-to-infrared generation models. To establish a diverse and reliable training corpus, we aggregate 69 publicly available infrared datasets and sources.

The collected sources cover diverse infrared scenarios, including ground-level perception, aerial and maritime imaging, surveillance and tracking, small-target detection, adverse-condition imaging, multispectral perception, and human-centered thermal imagery. Before preprocessing, these sources contain approximately 2.22 million image files or video frames, providing broad coverage of infrared scenes and appearance variations.

Rather than directly merging the original datasets, we construct a unified processing pipeline to extract valid infrared observations, standardize image representations, reduce temporal redundancy, and remove low-quality or duplicate samples. After preprocessing and quality control, the final corpus contains 1,107,778 infrared images drawn from all 69 collected sources. This curated corpus is used to train our text-to-infrared generation framework.

A.2 CORPUS CONSTRUCTION AND QUALITY CONTROL

The collected datasets exhibit substantial variations in modality organization, image representation, and annotation formats, including paired visible–thermal images, multispectral observations, video sequences, and auxiliary annotations. We therefore perform source-aware preprocessing to extract valid infrared observations, remove non-infrared auxiliary data, and unify image representations across different sources. For high-bit-depth and floating-point thermal images, we preserve their native numerical representations during preprocessing to avoid altering the original infrared intensity distributions and maintain physically meaningful thermal patterns.

Since many infrared datasets are derived from video sequences, directly retaining all frames may introduce severe temporal redundancy and bias the training distribution toward a small number of long sequences. We therefore apply source-dependent temporal sampling to video-based datasets while keeping independent image collections unchanged. After modality extraction, representation standardization, and temporal sampling, we obtain 1,253,393 standardized infrared candidates.

To further improve corpus reliability, we perform a multi-stage quality control procedure involving validity filtering, exact-content deduplication, perceptual duplicate analysis, and semantic quality screening. Exact duplicates are removed through SHA-256 verification, while perceptually similar samples are considered during dataset partitioning to prevent potential leakage. Finally, a vision-language model is employed to identify and remove samples unsuitable for image generation, producing the final curated infrared corpus.

A.3 SEMANTIC AND RADIOMETRIC ANNOTATION

Existing infrared datasets generally lack unified textual descriptions, which limits their direct use for text-conditioned generation. To construct image-text pairs for infrared synthesis, we generate structured semantic annotations for all retained infrared images using Qwen2.5-VL with a structured prompting strategy. The generated descriptions characterize scene content, primary objects, spatial relationships, viewpoint, and other visually observable properties, providing semantic guidance for text-to-infrared generation.

To improve annotation reliability, the captioning process is designed to describe only information supported by visual evidence and avoid overly specific semantic assumptions under ambiguous conditions. We generate captions for all 1,107,778 retained images and perform post-generation verification to ensure complete annotation coverage.

Beyond semantic descriptions, we further extract image-derived radiometric attributes to characterize infrared appearance. Specifically, we compute statistics related to brightness, contrast, dynamic

range, and local intensity variations, and discretize them into several levels based on data-driven quantiles. These attributes provide coarse descriptions of infrared appearance and complement semantic captions with additional appearance-level guidance.

The radiometric attributes are treated as visual characteristics of infrared images rather than physical sensor measurements. They do not represent calibrated temperature, sensor noise, or specific infrared waveband information, which remain unspecified when reliable metadata are unavailable.

A.4 Leakage-Safe Dataset Split

Random image-level splitting is unreliable for large-scale infrared datasets because many samples originate from video sequences or are repeatedly collected across different benchmarks. Such splitting may introduce highly similar observations into both training and evaluation sets, resulting in overly optimistic performance estimates.

To prevent potential leakage, we construct training, validation, and test partitions based on connected components rather than individual images. Samples are grouped when they share the same source sequence, identical SHA-256 content identity, or membership in the same perceptual near-duplicate group. All samples within the same group are assigned to a single partition, ensuring that closely related observations do not appear across different splits.

For sources with reliable temporal information, sequence-level grouping is applied before partitioning. Sources without reliable sequence structures are conservatively assigned to the training set. The validation and test sets are further constructed using source-aware sampling to preserve diversity across infrared scenarios while avoiding domination by large-scale datasets.

The final split contains 1,092,778 training images, 5,000 validation images, and 10,000 test images. A final integrity check confirms no overlap among different partitions with respect to sequence groups, duplicate components, and available image identities. Validation5K is used for model selection, while Test10K is reserved for final evaluation.

A.5 Corpus Summary and Limitations

The final corpus contains 1,107,778 infrared images collected from 69 publicly available infrared datasets and sources, including 1,092,778 training images, 5,000 validation images, and 10,000 test images. The collected data cover diverse infrared applications, including RGB-T perception, object tracking, small-target detection, aerial imaging, and multispectral thermal analysis.

Since the corpus is aggregated from multiple public infrared datasets, it naturally inherits several limitations from existing benchmarks. The contributions of individual sources are imbalanced, and sensor-related metadata are incomplete for some datasets. In addition, the displayed infrared intensity values may vary across different acquisition devices and imaging conditions and should not be interpreted as calibrated physical temperature measurements.

Although our preprocessing pipeline reduces redundancy, removes invalid samples, and improves evaluation reliability, it cannot completely eliminate biases inherited from existing datasets. Therefore, the proposed corpus is designed for large-scale infrared appearance modeling and text-conditioned generation rather than physical temperature estimation or sensor-specific analysis.

A.6 Collected Dataset Sources

The constructed infrared corpus is aggregated from 69 publicly available infrared datasets and sources. Since these datasets were originally collected for different infrared research tasks, we organize them according to their primary research purposes rather than strict scene categories. The collected sources cover major infrared research areas, including RGB-T perception, object tracking, infrared small-target detection, multispectral imaging, aerial perception, and application-specific thermal analysis.

RGB-T perception and paired infrared-visual datasets. The corpus includes representative RGB-T perception datasets, including KAIST Hwang et al. (2015), FLIR-IISR Zou et al. (2026a), FLIR-IVSR Zou et al. (2026b), FLIR ADAS v2 FLIR Systems (2018), LLVIP Jia et al. (2021),

CVC-14 González et al. (2016), IVSD Jiang et al. (2024), RoadScene Xu et al. (2020), SODA Li et al. (2021), and TIDS Botsoglou et al. (2025). These datasets mainly provide paired visible-infrared observations for perception tasks such as pedestrian detection, autonomous driving, and scene understanding.

Tracking and surveillance. The collected sources also include tracking-oriented infrared datasets, such as LasHeR Li et al. (2022), RGBT210 Li et al. (2017), VT-MOT Zhu et al. (2025), VT5000 Tu et al. (2023), VT1000 Tu et al. (2020), VT821 Tu et al. (2020), TMTV Tran et al. (2026), TRI-CLOBS Toet et al. (2016), LTIR Berg et al. (2015), HDO Xie et al. (2024), JU-VNT Xie et al. (2024), Kust4K Ouyang et al. (2025), and UTV20K Wang et al. (2025a). These datasets contain diverse thermal observations collected for object tracking, surveillance, and long-term object tracking, covering various targets, environments, and challenging imaging conditions.

Infrared small-target detection and adverse conditions. For challenging infrared target detection scenarios, we incorporate small-target and adverse-condition datasets, including IRSTD-1k Zhang et al. (2022), SIRST-V2 Dai et al. (2023), IR-TD Cao et al. (2025), ISOD Zhang & Cai (2023), NTLNP Tan et al. (2022), SMOD Chen et al. (2025b), TBRSD Chen & Bai (2023), AAU-RainSnow Bahnsen & Moeslund (2019), and the turbulence-distorted infrared imaging dataset Wang (2023). These sources provide diverse target scales and challenging imaging conditions.

Multispectral and thermal imaging. The corpus further includes multispectral and thermal imaging datasets, including FMB Liu et al. (2023), M3FD Liu et al. (2022), MSRS Tang et al. (2022), MFNet Ha et al. (2017), BU-TIV Wu et al. (2014), EMS Yi et al. (2024), MCubeS Liang et al. (2022), MM5 Brenner et al. (2026), MSVS Benezeth et al. (2014), NII-CU Speth et al. (2022), PST900 RGBT Shivakumar et al. (2020), R-LiViT Mirlach et al. (2025), RGBNIR Brown & Süsstrunk (2011), TIR-SS Chen et al. (2022), TNO Toet (2014), and TRISTAR RGB-Thermal-Depth Stippel et al. (2024). These datasets provide complementary infrared observations collected from multispectral sensing platforms.

Aerial and application-specific thermal datasets. Finally, we include aerial and application-specific thermal datasets, such as Anti-UAV-RGBT Jiang et al. (2021), AVIID Han et al. (2023), BIRDSAI Bondi et al. (2020), Caltech Aerial RGB-Thermal Lee et al. (2024), DroneRGBT Peng et al. (2021), HIT-UAV Suo et al. (2023), RGB-LWIR air-based Gallagher & Oughton (2023), Tri-band Vehicle and Vessel Liu et al. (2025), UAV_RGB-T_2400 Song et al. (2023), VAIS Zhang et al. (2015), VeDAI Razakarivony & Jurie (2016), Gas-DB Wang et al. (2024), Thermal Anomaly Segmentation Ruck et al. (2025), and other infrared sources.

B Implementation Details

B.1 Training Configuration

Backbone and optimization. All experiments are conducted using Stable Diffusion v1.5 (SD1.5) at a resolution of 512×512. We use the pretrained AutoencoderKL and CLIP text encoder from SD1.5 while retaining its original four-channel latent interface. In Stage I, the original VAE encoder, posterior head, and decoder are frozen, and only the Frequency-Split Adapters (FSAs) and Latent Mean Refiner are optimized. The FSAs are inserted into three encoder blocks with a bottleneck width of 32, while the Latent Mean Refiner operates on the four-channel posterior mean with a bottleneck width of 16. In Stage II, the complete VAE, IR-FAE modules, and text encoder are frozen, while only the denoising U-Net is fine-tuned. The latent scaling factor follows the original SD1.5 configuration with $\gamma = 0.18215$.

All models are trained using BF16 mixed precision. Stage I is optimized using AdamW with the corresponding learning rate setting, while Stage II uses AdamW with a learning rate of 5×10^{-6}, cosine learning-rate decay, and 150 warmup steps. The effective batch size is set to 32, and the U-Net is optimized for 5,000 steps on NVIDIA GeForce RTX 5090 GPUs. Model selection is performed on Validation5K, while Test10K is reserved for final evaluation.

Infrared preprocessing. Infrared images are converted into the input format required by the pretrained VAE. Single-channel infrared images are replicated across three channels, normalized to $[-1, 1]$, and resized to 512×512. During training, we apply random 512×512 cropping and horizontal flipping with probability 0.5 for data augmentation. Validation and reconstruction evaluation use deterministic resizing and center cropping.

Text conditioning. Stage-II training uses the automatically generated infrared captions described in Appendix A.3. The captions are encoded using the frozen SD1.5 CLIP text encoder, and no text-encoder parameters are updated during training.

B.2 Stage-I Optimization

Stage I learns an infrared-aware latent representation through frequency-aware feature adaptation and posterior-mean refinement. The training objective in Eq. (6) of the main paper is formulated as

$$\mathcal{L}_{\mathrm{Stage-I}} = \lambda_{\mathrm{img}}\mathcal{L}_{\mathrm{img}} + \lambda_{\mathrm{freq}}\mathcal{L}_{\mathrm{freq}} + \lambda_{\mathrm{lat}}\mathcal{L}_{\mathrm{lat}} + \lambda_{\mathrm{gate}}\mathcal{L}_{\mathrm{gate}}. \tag{9}$$

The image loss preserves reconstruction fidelity through pixel-level and structural constraints:

$$\mathcal{L}_{\mathrm{img}} = \|\widehat{x}_{\mathrm{IR}} - x\|_1 + \beta_{\mathrm{ssim}}(1 - \mathrm{SSIM}(\widehat{x}_{\mathrm{IR}}, x)). \tag{10}$$

The frequency loss encourages consistency between reconstructed and original infrared frequency characteristics. Specifically, it compares their frequency-domain representations and normalized energy distributions across different frequency bands:

$$\mathcal{L}_{\mathrm{freq}} = \|\log(1 + |\mathcal{F}(\widehat{x}_{\mathrm{IR}})|) - \log(1 + |\mathcal{F}(x)|)\|_1 + \beta_{\mathrm{band}}\mathcal{L}_{\mathrm{band}}. \tag{11}$$

The three frequency bands are defined according to normalized radial frequency boundaries of 0.15 and 0.35. The latent compatibility loss constrains the adapted posterior mean to remain close to the pretrained latent representation while preserving its statistical characteristics:

$$\mathcal{L}_{\mathrm{lat}} = \|\mu_{\mathrm{IR}} - \mu_{\mathrm{base}}\|_1 + \beta_{\mathrm{stat}}\mathcal{L}_{\mathrm{stat}} + \beta_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}. \tag{12}$$

The KL regularization provides a weak constraint on the adapted posterior while maintaining compatibility with the pretrained latent distribution. The gate regularization loss prevents the adaptive fusion module from collapsing to a single frequency branch. Only the FSAs and Latent Mean Refiner are optimized during this stage, while all original VAE parameters remain frozen.

B.3 Stage-II Progressive Latent Adaptation

During Stage II, the denoising U-Net is adapted using Progressive Latent Adaptation (PLA). At optimization step k, the adaptation strength is defined as

$$\rho_k = 0.5 \min\left(\frac{k}{1250}, 1\right), \tag{13}$$

and the latent mean is progressively interpolated by

$$\mu_k = (1 - \rho_k)\mu_{\mathrm{base}} + \rho_k\mu_{\mathrm{IR}}. \tag{14}$$

The clean latent used for diffusion training is sampled as

$$z_0^{(k)} = \gamma\left(\mu_k + \sigma_{\mathrm{base}} \odot \xi\right), \qquad \xi \sim \mathcal{N}(0, I), \tag{15}$$

where the posterior variance from the pretrained VAE is preserved. The sampled latent is perturbed using the original SD1.5 diffusion forward process, and the U-Net is optimized with the standard

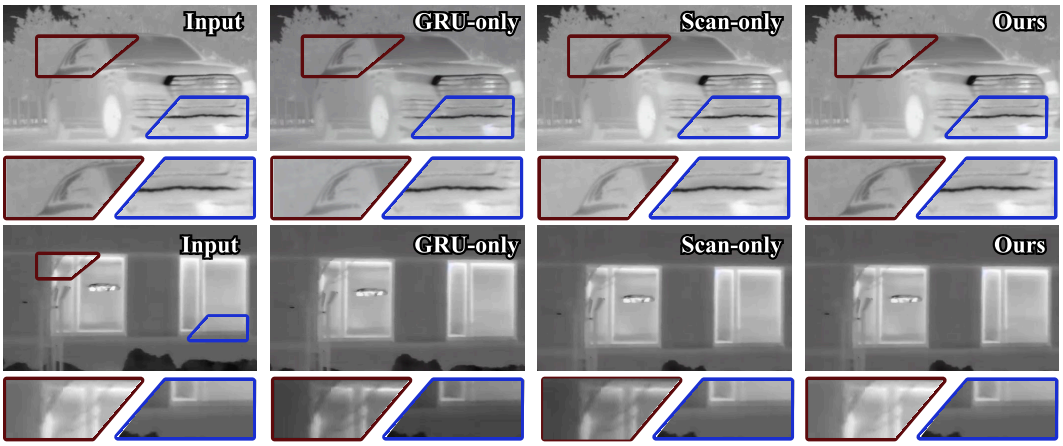


Figure 9: Qualitative comparison of different global context modeling strategies. Red and blue boxes highlight global thermal structures and local details, respectively.

noise-prediction objective. PLA modifies only the clean latent distribution during training and does not introduce additional inference-time modules.

B.4 Inference Configuration

All reported SD1.5 generation results are produced at 512×512 resolution using 30 denoising steps and classifier-free guidance with scale 7.5. All compared methods use the same prompt set, image resolution, and evaluation protocol. We generate one image for each Test10K prompt and use identical random seeds for fair comparison. No negative prompt is used during inference.

At inference time, neither PLA nor IR-FAE is executed. FALA uses the adapted U-Net together with the frozen text encoder, diffusion scheduler, and VAE decoder following the standard latent-diffusion pipeline. Therefore, FALA introduces no additional inference-time module, parameter, or denoising step compared with the corresponding SD1.5 full fine-tuning baseline.

C Additional Experiments

C.1 Effect of Global Context Modeling

The low-frequency branch of IR-FAE is designed to capture spatially coherent thermal structures by modeling long-range dependencies. To investigate the effectiveness of different global context modeling strategies, we compare the complete global context module with two simplified variants, including GRU-only modeling and selective scan-only modeling.

All variants are trained under the same configuration and evaluated on the same fixed validation subset. The Full model adopts the complete design that combines bidirectional recurrent aggregation and selective scanning, while the two variants remove one of the two global context components.

Table 2: Ablation of global context modeling in the low-frequency branch of IR-FAE.

Method	PSNR↑	SSIM↑	Spectral Error↓	Band Energy Error↓
Full IR-FAE	**38.4879**	**0.9487**	**0.004284**	**1.057**$\times 10^{-5}$
GRU-only	37.9398	0.9465	0.004448	1.317$\times 10^{-5}$
Selective Scan-only	37.9626	0.9465	0.004448	1.316$\times 10^{-5}$

Table 2 reports the quantitative comparison. The complete IR-FAE achieves the best performance across all evaluated metrics, obtaining a PSNR of 38.4879 and an SSIM of 0.9487, while also achieving the lowest spectral and band-energy errors. Removing either the GRU-based directional aggre-

gation or the selective scanning module consistently decreases reconstruction quality and increases frequency errors. These results indicate that the two global context components provide complementary information for modeling infrared thermal structures. The recurrent aggregation improves directional dependency modeling, while selective scanning further enhances long-range spatial interactions in complex thermal distributions.

Figure 9 provides qualitative comparisons between different global context modeling strategies. Compared with the single-component variants, the complete model better preserves large-scale thermal distributions and local intensity transitions. The highlighted regions show that combining recurrent aggregation and selective scanning produces more coherent infrared structures with improved thermal consistency.

C.2 Effect of Latent Mean Refinement and Fusion Regularization

IR-FAE adapts infrared latent representations through frequency-aware feature correction and latent-compatible posterior refinement. Branch dropout and gate regularization stabilize adaptive fusion between low-frequency thermal structures and high-frequency details.

To evaluate these components, we perform controlled ablations by removing the Latent Mean Refiner, branch dropout, and gate regularization. All variants are trained with the same training configuration and evaluated using the same validation protocol. As shown in Table 3, removing the Latent

Table 3: Ablation study of latent refinement and fusion regularization.

Method	PSNR↑	SSIM↑	Spectral↓	Band↓
Full	**38.3732**	**0.946686**	**4.4424**	**1.7310**
w/o Latent Mean Refiner	37.6749	0.946453	4.4528	1.7648
w/o Branch Dropout	38.0925	0.946583	4.4461	1.7497
w/o Gate Regularization	37.9003	0.946527	4.4491	1.7571

Mean Refiner causes a clear degradation in reconstruction fidelity and frequency consistency. The PSNR decreases from 38.3732 to 37.6749, while spectral error increases from 4.4424 to 4.4528 and band energy error increases from 1.7310×10^{-5} to 1.7648×10^{-5}. This indicates that adapting intermediate encoder features alone is insufficient, and an additional correction on the final four-channel latent mean is important for recovering infrared-specific latent structures.

Removing branch dropout or gate regularization also results in degraded structural and frequency consistency. Specifically, the removal of branch dropout reduces PSNR to 38.0925 and increases band energy error to 1.7497×10^{-5}, while removing gate regularization further increases frequency errors. These results indicate that both regularization strategies help maintain balanced interactions between different frequency branches and prevent unstable frequency fusion.

Although LPIPS and DISTS show relatively small variations among different variants, the improvements in PSNR, spectral error, and band energy error demonstrate that these components mainly contribute to more accurately preserving fine-grained infrared structure and physically meaningful frequency characteristics rather than general perceptual similarity.