CoVisIT: Cross-Modal Prior Guided Diffusion Model for Visible-to-Infrared Image Translation

Linfeng Tang[1,†] **Tong Hu**[2,†] **Zizhuo Li**[2] **Hao Zhang**[2]
Han Xu[3,*] **Zhenfeng Shao**[4] **Jiayi Ma**[1,2,*]

[1]School of Robotics, Wuhan University [2]Electronic Information School, Wuhan University
[3]School of Automation, Southeast University [4]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University
{linfeng0419, vod10311, jyma2010}@gmail.com, xu_han@seu.edu.cn

Abstract

Visible sensors provide rich scene details but are vulnerable to adverse conditions, whereas infrared sensors capture thermal radiation distributions and are more robust to environmental variations. However, acquiring high-resolution infrared images remains challenging due to sensor limits as well as hardware costs, and collecting pixel-wise aligned visible-infrared pairs is even more difficult. Existing methods mainly rely on infrared super-resolution or visible-to-infrared (VIS-to-IR) translation: the former preserves thermal distributions but may underutilize visible cues for fine-detail recovery and spatial alignment, while the latter benefits from fine-grained structural cues but faces an ill-posed thermal-distribution inference problem. To overcome these limitations, we propose CoVisIT, a cross-modal prior guided diffusion model for VIS-to-IR image translation. Our key insight is to constrain VIS-to-IR translation with complementary cross-modal priors, using high-resolution visible images to recover fine-grained scene details and misaligned low-resolution infrared inputs to anchor reliable thermal distributions. To achieve this, we build upon a latent diffusion framework and develop a Cross-Modal Adaptation module that modulates visible and infrared features in a shared feature space, thereby reducing the modality gap and exploiting complementary cross-modal priors. To further address local misalignment, we introduce a Dynamic Kernel Generation Module to predict input-adaptive kernels and embed a Multi-Scale Dynamic Convolution Module into the denoising U-Net, enabling dynamic local feature aggregation for implicit alignment. As a result, CoVisIT generates high-resolution, visible-aligned infrared images with reliable thermal distributions. Extensive experiments demonstrate state-of-the-art performance and consistent improvements in downstream tasks. The source code is publicly available at https://github.com/Linfeng-Tang/CoVisIT.

1 Introduction

In real-world scenarios, relying only on visible (VIS) sensors is often insufficient for robust visual perception. For example, scene information captured by the VIS modality can be severely degraded under low-light conditions, strong illumination changes, or occlusions. In contrast, infrared (IR) sensors provide sensing cues based on thermal radiation emitted by objects, which are usually more stable in complex environments. Therefore, IR images have become an important information source for downstream visual tasks, such as multi-modal image fusion [29, 20], object detection [9, 4],

[†] Equal contribution; [*] Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

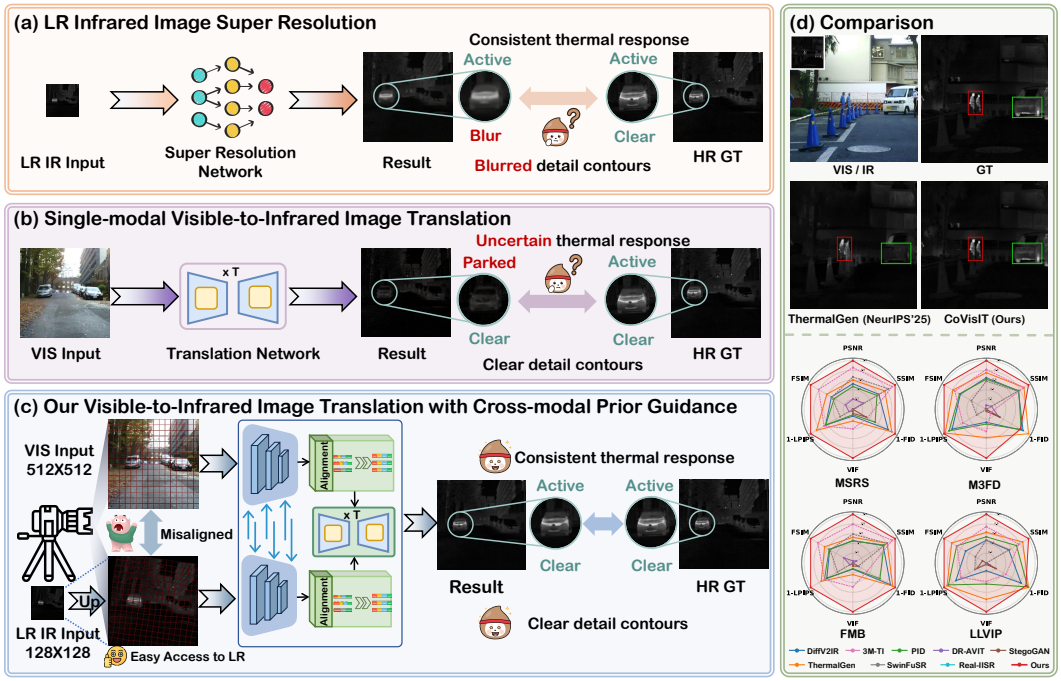


Figure 1: Existing super-resolution and translation paradigms suffer from blurred details or unreliable thermal responses. Using VIS texture cues and low-resolution IR thermal priors, our method generates VIS-aligned, high-resolution IR images with clearer structures and faithful thermal distributions.

and semantic segmentation [36, 21]. However, high-resolution (HR) IR image acquisition remains constrained by sensor resolution limits and acquisition costs. Moreover, heterogeneous sensor configurations often introduce non-negligible spatial misalignment between VIS and IR images, hindering the acquisition of pixel-aligned VIS-IR pairs. Consequently, the lack of HR IR images that are pixel-aligned with VIS images limits the development of related multi-modal vision tasks.

Extensive efforts have been devoted to obtaining HR IR images, and existing methods can be broadly categorized into two groups. The first is IR image super-resolution (SR) [16, 42], which reconstructs high-resolution (HR) IR images from low-resolution (LR) IR inputs, as shown in Fig. 1 (a). These methods preserve original IR thermal distributions, but often struggle to recover fine textural details, resulting in blurred contours and textures. To compensate for missing details, VIS-guided IR SR methods [40, 2] introduce VIS images as references. However, these methods are primarily designed to enhance IR resolution and often overlook the spatial misalignment between IR and VIS images, limiting their application in downstream multi-modal tasks. Furthermore, although reference-based IR SR methods exploit both IR thermal distributions and VIS references, the VIS modality is often used only for local detail guidance, limiting its effective contribution and potentially introducing texture artifacts. The other group is VIS-to-IR image translation [5, 23, 34], which directly generates IR images from VIS inputs, as shown in Fig. 1 (b). With the strong generation capabilities of generative models, this paradigm has become a common choice for obtaining HR IR images. Moreover, by using VIS images as the spatial reference, it can generate HR IR counterparts with the same field of view and scene layout as VIS inputs. However, these methods mainly infer IR content from VIS inputs and thus lack reliable thermal cues, often producing ambiguous or inaccurate thermal responses.

To address these challenges, we present a novel **C**ross-modal Prior Guided Diffusion Model for **Vis**ible-to-**I**nfrared **T**ranslation, termed **CoVisIT**. As shown in Fig. 1 (c), leveraging the powerful synthesis capacity of generative models, CoVisIT aims to generate detail-rich IR images using VIS images as texture references, while correcting thermal responses with IR distribution cues from LR IR inputs. Although VIS-guided IR SR and our IR-guided VIS-to-IR translation both take HR VIS and LR IR images as inputs and generate HR IR images, they differ in output geometry and modality roles. VIS-guided IR SR preserves the field of view and geometry of LR IR, whereas CoVisIT generates HR IR images aligned with HR VIS in both field of view and geometry. Specifically, HR VIS determines the output structure and texture, while LR IR provides thermal-distribution cues rather than serving as the geometric reference. To jointly introduce VIS texture priors and IR thermal-distribution

priors into the diffusion framework, we introduce a Cross-Modal Adaptation module to modulate VIS and IR features in a shared feature space, reducing modality representation gaps and enabling reliable cross-modal prior extraction. Given the inherent spatial misalignment between VIS and IR images, we embed a Multi-Scale Dynamic Convolution Module into the denoising U-Net, where input-adaptive kernels generated by the Dynamic Kernel Generation Module dynamically adjust local feature aggregation during denoising to achieve implicit alignment. In addition, auxiliary text cues extracted from the input IR image are used to better indicate the distribution of thermal objects. As shown in Fig. 1 (d), CoVisIT produces clearer structures and more faithful thermal distributions. Experiments on LLVIP, FMB, M3FD, and MSRS demonstrate the comprehensive advantages of the proposed method. In summary, our main contributions are as follows:

• We propose **CoVisIT**, a cross-modal prior guided diffusion model for visible-to-infrared translation that synergistically exploits HR VIS texture priors and misaligned LR IR thermal-distribution cues to generate VIS-aligned HR IR images with clear textures and faithful thermal responses.

• A Cross-Modal Adaptation module is designed to reduce VIS-IR representation gaps through bidirectional feature calibration. Local spatial misalignment is addressed through scale-aware IR feature reassembly, where dynamic kernels generated by the Dynamic Kernel Generation Module are used in the Multi-Scale Dynamic Convolution Module.

• Extensive experiments demonstrate that CoVisIT achieves superior visual quality, distribution fidelity, and perceptual quality, while bringing clear improvements to downstream multi-modal tasks.

2 Related work

Infrared image super resolution. Early IR SR works usually employ VIS SR paradigms [3, 17] for IR reconstruction [24, 41]. To better adapt to IR imaging characteristics, IR-specific methods have been explored. For example, DifIISR [16] injects gradient and perceptual priors into diffusion with thermal spectrum regulation, while Real-IISR [42] handles real-world degradations using thermal-structural guidance and an autoregressive framework. Despite these advances, single-modality IR SR methods still rely only on LR IR inputs, making it difficult to recover fine scene structures.

To compensate for missing structural details, VIS-guided IR SR methods introduce HR VIS images as guidance. CoReFusion [12] fuses VIS and IR features with contrastive regularization, while SwinFuSR [1] uses cross-domain attention for guided reconstruction. Diffusion-based methods such as DuaDiff [27] further use dual-conditional diffusion with multi-modal latent cross-attention. Recently, 3M-TI [2] handles weakly aligned inputs with cross-modal self-attention and misalignment augmentation. However, these methods mainly target IR resolution enhancement rather than generating VIS-aligned IR counterparts. Their limited spatial-misalignment modeling restricts the effective use of complementary cross-modal priors and may introduce misaligned VIS textures into the reconstructed IR images as artifacts.

Visible-to-infrared image translation. Different from SR-based paradigms, visible-to-infrared (VIS-to-IR) image translation directly synthesizes IR images from VIS inputs and has been widely explored for IR image generation. Early methods [30, 10, 14, 13, 15, 43, 7] mainly use GAN-based frameworks and additional constraints to improve content consistency. For example, InfraGAN [43] adopts structural similarity constraints, while DR-AVIT [7] introduces geometry-consistency constraints for aerial VIS-to-IR translation. Recently, stronger generative models have been explored. PID [22] introduces physics-informed constraints, DiffV2IR [23] combines diffusion models with vision-language understanding, and ThermalGen [34] improves robustness across sensors, viewpoints, and environments via style-disentangled generation. Despite these advances, existing translation methods still infer IR appearance mainly from VIS-centric priors and lack explicit thermal observations, making thermal distributions ambiguous.

3 Methodology

3.1 Overall framework

Given an HR VIS image $I_{vis}^{hr} \in \mathbb{R}^{H\times W\times 3}$, existing image translation methods leverage generative models to synthesize a detail-rich IR counterpart. However, VIS-to-IR translation is inherently ill-posed because VIS images lack explicit thermal observations, often causing inaccurate thermal

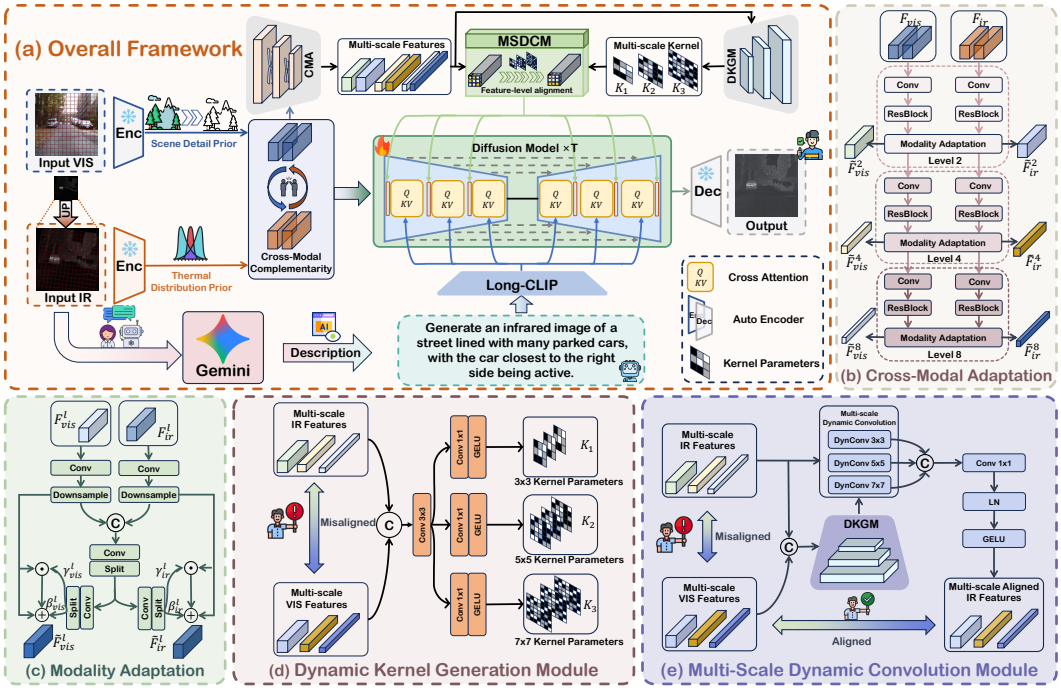


Figure 2: Overall framework of cross-modal prior guided diffusion model for VIS-to-IR translation.

responses. LR IR images $I_{ir}^{lr} \in \mathbb{R}^{h\times w\times 1}$, *which can be acquired at relatively low cost*, provide explicit thermal-distribution references despite their spatial misalignment with VIS images. Moreover, scene descriptions T_{ir} extracted from I_{ir}^{lr} by a vision-language large model (*e.g.,* Gemini) help the generative model understand thermal target distributions. Therefore, with I_{vis}^{hr}, I_{ir}^{lr}, and T_{ir} as inputs, our goal is to generate an HR, VIS-aligned IR image $\hat{I}_{ir}^{hr} \in \mathbb{R}^{H\times W\times 1}$.

Specifically, as shown in Fig. 2 (a), I_{ir}^{lr} is first upsampled to match the spatial resolution of I_{vis}^{hr}. A retrained autoencoder then maps I_{vis}^{hr} and the upsampled I_{ir}^{lr} into latent features $\mathbf{F}_{vis} = Enc(I_{vis}^{hr})$ and $\mathbf{F}_{ir} = Enc(Up(I_{ir}^{lr}))$, while Long-CLIP [35] encodes T_{ir} into a text condition $\mathbf{c}_{ir} = E_{\text{long-clip}}(T_{ir})$. To improve thermal fidelity while preserving rich scene details, CoVisIT uses $\mathbf{F}_{vis}$ as a scene-detail prior and introduces $\mathbf{F}_{ir}$ as an explicit thermal-distribution cue. To enable reliable cross-modal prior extraction, a Cross-Modal Adaptation module (CMA) is developed to modulate VIS and IR features in a shared feature space as $(\widetilde{\mathbf{F}}_{vis}, \widetilde{\mathbf{F}}_{ir}) = \mathcal{A}_{cma}(\mathbf{F}_{vis}, \mathbf{F}_{ir})$. Although the modality-adapted features reduce the representation gap, local spatial misalignment between VIS and IR cues may still remain. Therefore, DKGM predicts multi-scale input-adaptive kernels $\mathcal{K}_{dyn} = \mathcal{G}_{dkgm}(\widetilde{\mathbf{F}}_{vis}, \widetilde{\mathbf{F}}_{ir})$, and MSDCM uses them to reassemble IR features for implicit alignment, yielding the aligned IR feature $\widehat{\mathbf{F}}_{ir} = \mathcal{M}_{msdcm}(\widetilde{\mathbf{F}}_{ir}, \mathcal{K}_{dyn})$. Starting from a Gaussian noise feature $\mathbf{F}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the conditional denoising network $\boldsymbol{\epsilon}_\theta$ iteratively predicts the noise at each timestep conditioned on the noisy latent $\mathbf{F}_t$, the VIS texture feature $\mathbf{F}_{vis}$, the aligned IR thermal-distribution feature $\widehat{\mathbf{F}}_{ir}$, and the text condition $\mathbf{c}_{ir}$. Through the reverse denoising process, CoVisIT obtains the target HR IR latent feature $\hat{\mathbf{F}}_{ir}^{hr}$, which is finally decoded into the image space as $\hat{I}_{ir}^{hr} = \mathrm{Dec}(\hat{\mathbf{F}}_{ir}^{hr})$.

3.2 Network architecture

Cross-modal adaptation. VIS and IR inputs provide complementary scene-detail and thermal-distribution priors, respectively, but modality representation gaps hinder their direct interaction. To address this issue, we introduce a Cross-Modal Adaptation (CMA) module, which performs cross-modal context-driven bidirectional feature calibration, as shown in Fig. 2 (b). Based on the latent features $\mathbf{F}_{vis}$ and $\mathbf{F}_{ir}$ extracted by the autoencoder, we construct multi-scale feature representations at different U-Net levels. Here, $l \in \{1, 2, 4, 8\}$ denotes the downsampling level. At level l, the corresponding VIS and IR features are denoted as $\mathbf{F}_{vis}^{l}$ and $\mathbf{F}_{ir}^{l}$. We first obtain their base

representations through modality-specific mappings:

$$\mathbf{B}_{vis}^{l} = \phi_{vis}^{l}(\mathbf{F}_{vis}^{l}), \qquad \mathbf{B}_{ir}^{l} = \phi_{ir}^{l}(\mathbf{F}_{ir}^{l}), \tag{1}$$

where $\phi_{vis}^{l}(\cdot)$ and $\phi_{ir}^{l}(\cdot)$ denote feature mapping functions for the VIS and IR branches. $\mathbf{B}_{vis}^{l}$ and $\mathbf{B}_{ir}^{l}$ are the corresponding outputs, which are jointly processed to extract cross-modal context:

$$\mathbf{C}^{l} = \psi^{l}\left(\mathbf{B}_{vis}^{l}, \mathbf{B}_{ir}^{l}\right), \tag{2}$$

where $\psi^{l}(\cdot)$ is the cross-modal context modeling function, and $\mathbf{C}^{l}$ denotes the joint context feature. As shown in Fig. 2 (c), instead of directly using $\mathbf{C}^{l}$ as the fused result, we further use it to predict distribution calibration parameters for both modalities, formulated as:

$$[\boldsymbol{\gamma}_{vis}^{l}, \boldsymbol{\beta}_{vis}^{l}, \boldsymbol{\gamma}_{ir}^{l}, \boldsymbol{\beta}_{ir}^{l}] = \eta^{l}(\mathbf{C}^{l}), \tag{3}$$

in which $\eta^{l}(\cdot)$ denotes the parameter generation network, whose output is partitioned along the channel dimension by the Split operation in Fig. 2(c) to obtain modality-specific scale ($\boldsymbol{\gamma}$) and shift ($\boldsymbol{\beta}$) parameters. Finally, the two modality features are adaptively calibrated by affine modulation:

$$\widetilde{\mathbf{F}}_{vis}^{l} = \mathbf{B}_{vis}^{l} \odot (1 + \boldsymbol{\gamma}_{vis}^{l}) + \boldsymbol{\beta}_{vis}^{l}, \qquad \widetilde{\mathbf{F}}_{ir}^{l} = \mathbf{B}_{ir}^{l} \odot (1 + \boldsymbol{\gamma}_{ir}^{l}) + \boldsymbol{\beta}_{ir}^{l}, \tag{4}$$

where $\odot$ denotes element-wise multiplication, and $\widetilde{\mathbf{F}}_{vis}^{l}$ and $\widetilde{\mathbf{F}}_{ir}^{l}$ are the calibrated VIS and IR features. Through context-driven bidirectional calibration, CMA converts cross-modal complementarity into adaptive feature distribution adjustment, allowing each branch to absorb complementary cues while preserving modality-specific characteristics. This reduces representation gaps and provides consistent multi-scale cross-modal features for subsequent dynamic kernel generation and implicit alignment.

Dynamic kernel generation module. Although CMA reduces modality representation gaps, the modulated VIS and IR features may still suffer from local spatial misalignment. Notably, in practical VIS-IR imaging systems, hardware calibration typically mitigates large cross-modal displacement, while the remaining misalignment is mainly local and spatially variant. Therefore, we introduce a Dynamic Kernel Generation Module (DKGM) to produce sample-adaptive kernels for implicit alignment, as shown in Fig. 2 (d). Instead of relying on global geometric modeling, DKGM learns alignment-aware dynamic kernels from VIS-IR context. These kernels act as spatially adaptive local operators that reassemble IR features around each location under VIS structural guidance, converting local cross-modal displacement into learnable feature aggregation. By generating kernels of different sizes, DKGM covers misalignment across multiple spatial ranges and provides effective dynamic kernels for subsequent IR feature alignment.

Specifically, at the l-th scale, DKGM generates a level-specific dynamic kernel set $\mathcal{K}_{dyn}^{l} = \{\mathbf{K}_{m}^{l}\}_{m=1}^{M_l}$ from the modulated VIS feature $\widetilde{\mathbf{F}}_{vis}^{l}$ and IR feature $\widetilde{\mathbf{F}}_{ir}^{l}$, where M_l denotes the number of kernel branches at level l. For the m-th branch with kernel size k_m^l, DKGM first predicts a group-shared kernel weight through the dynamic kernel generation branch $\varphi_{m}^{l}(\cdot)$:

$$\mathbf{W}_{m}^{l} = \varphi_{m}^{l}\left(\widetilde{\mathbf{F}}_{vis}^{l}, \widetilde{\mathbf{F}}_{ir}^{l}\right) \in \mathbb{R}^{G \times (k_m^l)^2}, \tag{5}$$

where G denotes the number of channel groups. The predicted weights are then normalized and reshaped into group-shared dynamic kernels:

$$\widehat{\mathbf{K}}_{m}^{l} = \mathrm{Reshape}\big(\mathrm{Norm}\big(\mathbf{W}_{m}^{l}\big)\big) \in \mathbb{R}^{G \times k_m^l \times k_m^l}, \tag{6}$$

$$\mathbf{K}_{m}^{l} = \mathrm{Repeat}_{C_g}\left(\widehat{\mathbf{K}}_{m}^{l}\right) \in \mathbb{R}^{C_l \times k_m^l \times k_m^l}, \tag{7}$$

where C_l denotes the number of channels at the current scale, $C_g = C_l/G$ is the group channel size, and $\mathrm{Repeat}_{C_g}(\cdot)$ replicates each group-shared kernel across its C_g feature channels. Compared with fully channel-wise kernel generation, the group-shared design substantially reduces dynamic kernel parameters, alleviating training instability and memory overhead caused by high-dimensional kernels. The generated dynamic kernels provide different receptive fields for local alignment, where small kernels capture fine local deviations and larger kernels handle relatively larger spatial offsets. As a result, DKGM offers stable input-adaptive operators for subsequent implicit alignment, enabling more effective local misalignment recovery than fixed convolution kernels.

Table 1: Quantitative comparison of HR infrared image generation results. The best and second-best results are highlighted in red and purple, respectively.

Method	LLVIP						FMB					
	PSNR↑	SSIM↑	FSIM↑	VIF↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	FSIM↑	VIF↑	LPIPS↓	FID↓
Real-IISR [42]	25.512	0.780	0.874	0.168	0.215	68.486	31.092	0.917	0.934	0.214	0.150	70.508
SwinFuSR [1]	19.997	0.645	0.723	0.028	0.491	176.819	24.862	0.849	0.830	0.035	0.374	152.076
3M-TI [2]	24.707	0.765	0.854	0.142	0.363	124.324	29.281	0.903	0.907	0.158	0.256	131.705
StegoGAN [33]	11.893	0.340	0.672	0.032	0.518	144.971	12.132	0.486	0.690	0.051	0.514	150.551
DR-AVIT [7]	11.963	0.342	0.644	0.022	0.629	163.915	13.121	0.524	0.746	0.043	0.487	164.308
PID [22]	21.127	0.733	0.858	0.159	0.171	40.518	21.890	0.815	0.836	0.064	0.255	102.329
DiffV2IR [23]	19.889	0.634	0.793	0.072	0.260	62.450	22.066	0.806	0.833	0.067	0.276	78.088
ThermalGen [34]	22.725	0.814	0.893	0.259	0.153	**37.841**	23.490	0.848	0.864	0.095	0.226	72.813
CoVisIT (Ours)	**29.872**	**0.906**	**0.956**	**0.372**	**0.110**	43.437	**34.550**	**0.942**	**0.959**	**0.310**	**0.118**	**59.760**

Method	M3FD						MSRS					
	PSNR↑	SSIM↑	FSIM↑	VIF↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	FSIM↑	VIF↑	LPIPS↓	FID↓
Real-IISR [42]	28.597	0.853	0.902	0.177	0.191	54.402	32.853	0.887	0.920	0.163	0.232	50.770
SwinFuSR [1]	22.597	0.769	0.784	0.031	0.427	126.673	27.357	0.828	0.814	0.027	0.465	132.466
3M-TI [2]	27.320	0.839	0.876	0.133	0.307	116.291	31.750	0.877	0.898	0.121	0.316	104.821
StegoGAN [33]	11.127	0.416	0.701	0.040	0.510	97.419	12.318	0.368	0.681	0.047	0.583	107.003
DR-AVIT [7]	12.803	0.475	0.706	0.033	0.509	100.633	17.117	0.507	0.728	0.036	0.546	133.622
PID [22]	23.354	0.781	0.850	0.099	0.188	51.486	22.950	0.680	0.799	0.053	0.330	89.827
DiffV2IR [23]	23.977	0.777	0.845	0.091	0.214	47.550	24.231	0.704	0.817	0.060	0.330	64.144
ThermalGen [34]	26.058	0.840	0.886	0.161	0.154	**35.630**	25.946	0.781	0.846	0.083	0.246	57.022
CoVisIT (Ours)	**31.684**	**0.902**	**0.942**	**0.277**	**0.138**	50.368	**34.335**	**0.914**	**0.951**	**0.259**	**0.213**	**49.832**

Multi-scale dynamic convolution module. Based on the dynamic kernels generated by DKGM, we further design a Multi-Scale Dynamic Convolution Module (MSDCM) and connect it to different levels of the denoising U-Net for scale-aware implicit alignment, as shown in Fig. 2 (e). Since U-Net features across levels have distinct spatial scales and receptive fields, the dynamic kernels used for local alignment should also vary across levels. Therefore, we use kernel sizes $\{3,5,7\}$, $\{3,5\}$, and $\{3\}$ at levels 2, 4, and 8, respectively. Given the modulated IR feature $\widetilde{\mathbf{F}}_{ir}^{l}$, MSDCM applies each kernel in the corresponding dynamic kernel set $\mathcal{K}_{dyn}^{l} = \{\mathbf{K}_{m}^{l}\}_{m=1}^{M_l}$ via dynamic depthwise convolution and then aggregates the multi-branch responses:

$$\mathbf{Y}_{m}^{l} = \mathrm{DynConv}\left(\widetilde{\mathbf{F}}_{ir}^{l}, \mathbf{K}_{m}^{l}\right), \quad \mathbf{U}^{l} = [\mathbf{Y}_{1}^{l}, \ldots, \mathbf{Y}_{M_l}^{l}], \quad \mathbf{V}^{l} = \xi^{l}(\mathbf{U}^{l}), \tag{8}$$

where $\xi^{l}(\cdot)$ denotes the 1×1 fusion mapping. The aggregated feature is then refined by normalization, non-linear activation, and channel attention, and the aligned IR feature is obtained as:

$$\widehat{\mathbf{F}}_{ir}^{l} = \mathrm{CAM}\big(\sigma\big(\mathrm{Norm}\big(\mathbf{V}^{l}\big)\big)\big), \tag{9}$$

where $\mathrm{CAM}(\cdot)$ is the channel attention module. Finally, $\widehat{\mathbf{F}}_{ir}^{l}$ is injected into the corresponding U-Net level. With level-specific dynamic kernels, MSDCM performs scale-aware local IR feature reassembly under VIS guidance, enabling effective correction of spatially variant VIS-IR misalignment.

3.3 Loss functions

We optimize CoVisIT with three loss terms, including the diffusion loss, image reconstruction loss, and thermal-distribution consistency loss. The diffusion loss constrains the model to predict noise under multi-condition guidance:

$$\mathcal{L}_{diff} = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta}(\mathbf{F}_{t}, t, \mathbf{F}_{vis}, \widehat{\mathbf{F}}_{ir}, \mathbf{c}_{ir})\right\|_{2}^{2}\right], \tag{10}$$

where $\mathbf{F}_t$ denotes the noisy target IR latent feature at timestep t, $\boldsymbol{\epsilon}$ is the sampled Gaussian noise, and $\boldsymbol{\epsilon}_{\theta}(\cdot)$ is the noise prediction network. Considering that latent-space diffusion supervision alone cannot explicitly constrain image-level structures or thermal response distributions, we further impose image-space supervision. Specifically, the clean HR IR latent feature is estimated from the predicted noise following the standard denoising parameterization:

$$\hat{\mathbf{F}}_{ir}^{hr} = \left(\mathbf{F}_{t} - \sqrt{1-\bar{\alpha}_{t}}\boldsymbol{\epsilon}_{\theta}(\mathbf{F}_{t}, t, \mathbf{F}_{vis}, \widehat{\mathbf{F}}_{ir}, \mathbf{c}_{ir})\right)/\sqrt{\bar{\alpha}_{t}}, \tag{11}$$

which is decoded into an estimated HR IR image $\hat{I}_{ir}^{hr} = Dec(\hat{\mathbf{F}}_{ir}^{hr})$ for computing image-space losses. For VIS-to-IR image translation, local content accuracy and faithful thermal responses are both important. We first use an image reconstruction loss to preserve local appearance and edge structures between the generated image $\hat{I}_{ir}^{hr}$ and real HR IR image I_{ir}^{hr}. However, image-level supervision alone may over-smooth thermal regions or shift global thermal distributions. Therefore, we further introduce a thermal-distribution consistency loss to constrain global thermal response statistics:

$$\mathcal{L}_{rec} = \left\| \hat{I}_{ir}^{hr} - I_{ir}^{hr} \right\|_1 + \left\| \nabla \hat{I}_{ir}^{hr} - \nabla I_{ir}^{hr} \right\|_1, \qquad \mathcal{L}_{td} = \left\| \mathcal{H}(\hat{I}_{ir}^{hr}) - \mathcal{H}(I_{ir}^{hr}) \right\|_1, \tag{12}$$

where $\nabla(\cdot)$ denotes the spatial gradient operator, and $\mathcal{H}(\cdot)$ denotes the normalized soft thermal-response distribution operator. Finally, the total objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{diff} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{td}\mathcal{L}_{td}, \tag{13}$$

where $\lambda_{rec} = 1000$ and $\lambda_{td} = 10$.

4 Experiments

4.1 Implementation details and experiment configurations

Our method is built upon the latent diffusion framework [25]. A retrained autoencoder maps VIS/IR images into the latent feature space. The denoising backbone is a U-Net conditioned on the VIS texture feature, aligned IR thermal-distribution feature, and text condition. It uses downsampling scales $\{1, 2, 4, 8\}$ with channel multipliers $[1, 2, 4, 4]$, and applies cross attention at scales 2 and 4. CMA, DKGM, and MSDCM are embedded into the denoising U-Net to reduce modality gaps and handle local misalignment. We train CoVisIT on a mixed set of LLVIP [10], FMB [19], M3FD [18], and MSRS [28]. Input images are randomly cropped into 512×512 patches and augmented with random flipping and rotation. Scene descriptions are pre-extracted from LR IR inputs using the Gemini API. CoVisIT is trained with Adam (learning rate 1×10^{-5}, batch size 6, 1000 linear-schedule diffusion steps) on 4 NVIDIA A100 GPUs for 4 days, and uses 50-step DDIM sampling for inference.

Compared methods include IR SR method Real-IISR [42], VIS-guided IR SR methods SwinFuSR [1] and 3M-TI [2], and VIS-to-IR translation methods StegoGAN [33], DR-AVIT [7], PID [22], DiffV2IR [23], and ThermalGen [34]. For a fair comparison, all competing methods are retrained on the same mixed training set with their original model configurations and evaluated using identical data splits and evaluation metrics. For methods requiring LR IR inputs, the resolution is uniformly set to 128×128 with the same degradation process, while identical misalignment settings are applied to methods using both VIS and LR IR. PSNR [31], SSIM [32], FSIM [37], VIF [26], LPIPS [38], and FID [8] are used to quantify IR generation performance. All experiments are conducted on NVIDIA A100 GPUs and AMD EPYC 7543 Milan 2.8GHz CPUs with PyTorch.

4.2 Infrared image generation performance comparison

Quantitative comparison. Table 1 reports the quantitative results on LLVIP, FMB, M3FD, and MSRS. Real-IISR obtains competitive results by preserving LR IR thermal cues, but its outputs remain misaligned with VIS images. VIS-to-IR translation methods such as ThermalGen, DiffV2IR, and PID achieve competitive FID scores, but without IR distribution references, they struggle to anchor faithful thermal responses and lag behind on fidelity-oriented metrics. In contrast, CoVisIT synergistically exploits VIS texture priors and IR thermal priors, where CMA enables reliable prior extraction and MSDCM performs input-adaptive local alignment, thereby achieving the best performance on most metrics across the four datasets. Although VIS-guided IR SR methods such as SwinFuSR and 3M-TI also leverage both modalities, their limited misalignment handling prevents them from fully utilizing complementary priors. Overall, these quantitative results demonstrate that CoVisIT better balances structural recovery, VIS alignment, perceptual quality, and thermal-response fidelity.

Qualitative comparison. The visual results in Fig. 3 further support the above quantitative analysis. Specifically, in the low-light scenarios from LLVIP, VIS-to-IR translation methods such as PID, DiffV2IR, and ThermalGen fail to recover reliable thermal responses, showing the ambiguity of thermal inference from VIS-only inputs. Real-IISR preserves IR thermal cues but produces blurry details, while 3M-TI still suffers from detail degradation under modality gaps and spatial misalignment.

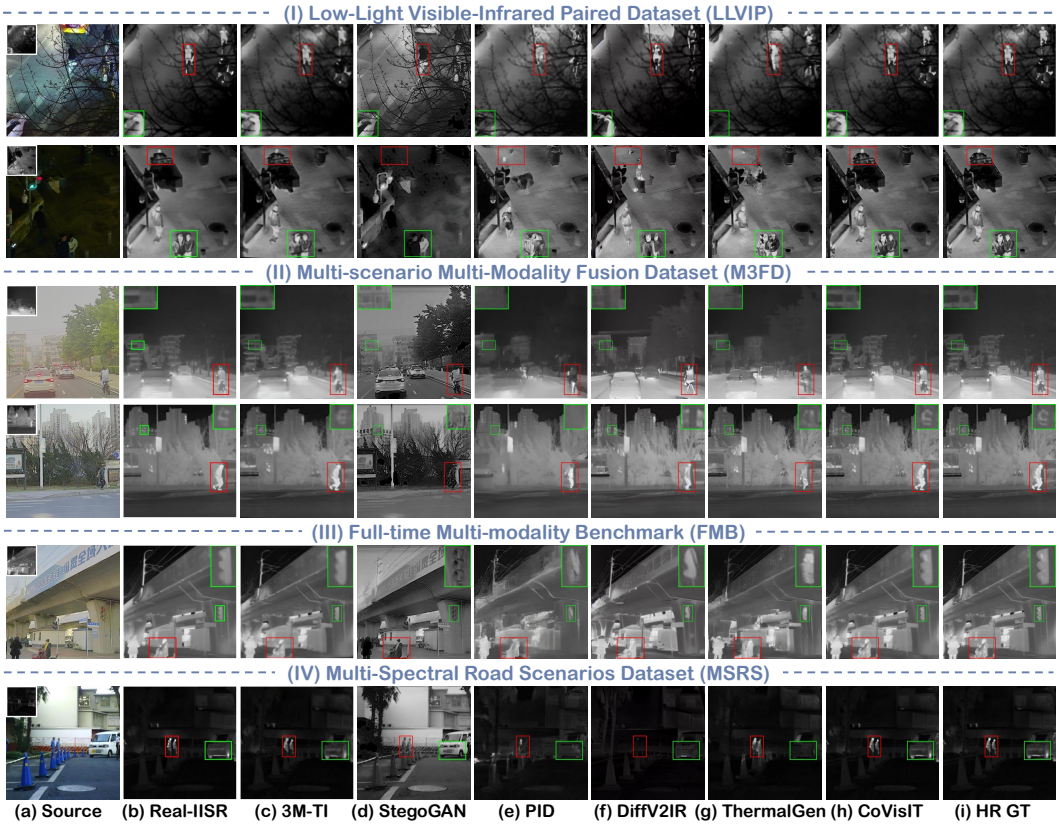


Figure 3: Qualitative comparison of high-resolution infrared image generation results.

Similar trends can be observed on FMB, M3FD, and MSRS, where complex road and urban scenes require both clear structures and stable thermal responses. In contrast, CoVisIT produces clearer contours, finer local details, and more faithful thermal responses, demonstrating the effectiveness of jointly using VIS texture priors, LR IR thermal-distribution cues, and input-adaptive local alignment.

4.3 Extended experiments

Robustness to deformation degrees. Real VIS-IR systems may retain local misalignment after hardware calibration, degrading cross-modal prior utilization and VIS-aligned IR generation. We therefore test robustness to nonrigid misalignment by applying random deformation flows with increasing elastic strengths to the IR input, while fixing rotation, translation, and Gaussian smoothing scale, as shown in Fig. 4 (e). Unlike 3M-TI that relies on misalignment augmentation and global cross-modal self-attention, CoVisIT handles misalignment with input-adaptive kernels that reassemble local IR features under VIS guidance. As deformation increases, 3M-TI shows blurred boundaries and local structural distortions, whereas CoVisIT preserves clearer and more accurate contours. Further experiments under diverse rotation, translation, and elastic deformation settings are provided in Appendix A.4.

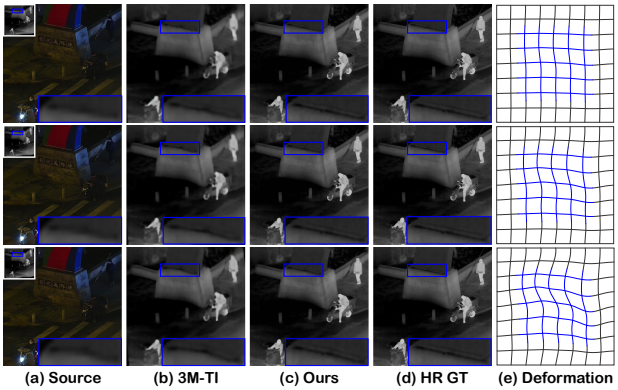


Figure 4: Comparison under varying deformation.

Image fusion. High-quality image fusion requires pixel-wise alignment and faithful thermal responses. Therefore, we employ a representative image fusion model, EMMA [39], to fuse IR images

Table 2: Semantic segmentation performance on the IR images generated by different methods.

Method	Road		Building		T-Sign		Veget.		Person		Car		mAcc	mIoU	aAcc
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU			
Real-IISR [42]	94.7	31.9	60.5	45.5	29.2	21.3	56.1	37.6	17.3	11.7	33.0	15.5	29.5	17.5	63.1
SwinFuSR [1]	78.1	68.8	85.2	59.5	49.6	46.0	73.1	59.7	27.9	26.2	67.2	52.8	40.5	33.7	79.2
3M-TI [2]	88.7	80.9	**90.8**	67.6	61.6	59.4	84.3	72.3	43.5	40.7	81.4	63.9	49.0	42.4	85.0
StegoGAN [33]	91.2	69.9	87.3	45.2	1.5	1.3	81.3	68.6	15.6	10.4	87.0	67.0	39.7	27.1	70.1
DR-AVIT [7]	74.6	62.1	78.1	38.7	0.4	0.4	38.7	33.5	4.9	3.7	57.8	44.5	29.2	20.8	67.2
PID [22]	93.2	83.2	81.3	69.1	33.8	32.8	88.3	71.1	47.8	41.0	83.0	64.3	46.5	37.4	85.9
DiffV2IR [23]	93.3	84.2	86.8	68.3	40.3	38.2	84.9	73.1	55.5	43.9	88.6	71.1	50.2	40.0	86.1
ThermalGen [34]	95.6	86.4	84.7	74.4	47.0	46.0	89.8	75.7	60.1	52.3	85.7	69.1	50.9	41.8	88.5
CoVisIT (Ours)	**95.7**	**88.5**	88.8	**79.1**	**65.3**	**62.7**	**90.1**	**78.4**	**68.3**	**63.2**	**91.2**	**74.3**	**58.6**	**48.2**	**90.1**
HR GT	96.9	89.8	90.9	82.0	71.3	68.9	91.0	81.6	76.7	70.1	93.3	77.5	62.6	52.0	91.4

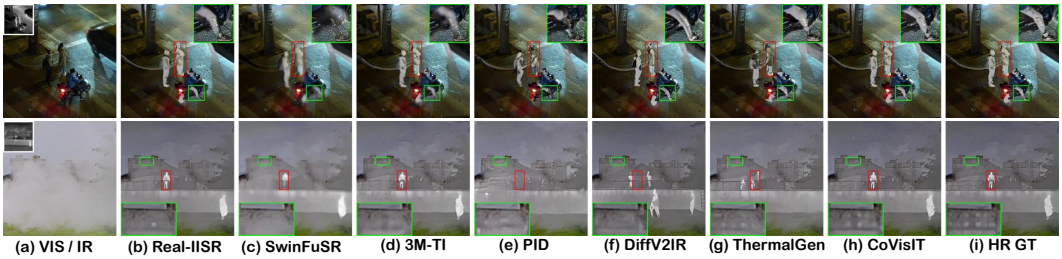


Figure 5: Fusion results based on IR images generated by different methods.

generated by different methods with their corresponding VIS images, further evaluating IR generation quality. As shown in Fig. 5, alternative methods often result in misaligned structures, blurred details, or inaccurate thermal targets, while CoVisIT provides better target saliency, clearer edges, and more consistent thermal responses. This demonstrates the high quality and downstream utility of CoVisIT-generated IR images. Further quantitative comparisons across the four datasets, together with detailed qualitative analyses of target saliency, structural preservation, and thermal-response consistency in the fusion results, are provided in the Appendix.

Semantic segmentation. Downstream perception tasks can further verify whether generated IR images contain useful semantic and thermal information. We first evaluate semantic segmentation performance of different methods on FMB, where SegNeXt [6] is retrained on ground-truth IR images and tested on images generated by different methods. As shown in Tab. 2, CoVisIT achieves the highest mIoU, mAcc, and aAcc, with clear gains on key foreground classes such as traffic sign, person, and car. This indicates that our generated images better preserve thermal cues and semantic structures for dense prediction. Qualitative results are provided in the Appendix.

Object detection. We further evaluate object detection performance on LLVIP to assess the downstream utility of the generated IR images. Specifically, we retrain YOLOv8 [11] on the ground-truth IR images from LLVIP as the detector. As shown in Tab. 3, our method achieves the best results across all metrics, indicating that the generated IR images preserve more discriminative thermal cues and object structures. Qualitative detection results are provided in the Appendix.

Table 3: Quantitative comparison of object detection.

Method	Prec.	Rec.	AP@0.50	AP@0.75	mAP@0.5:0.95
Real-IISR [42]	0.941	0.836	0.914	0.585	0.545
SwinFuSR [1]	0.653	0.360	0.358	0.033	0.109
3M-TI [2]	0.878	0.704	0.806	0.447	0.443
StegoGAN [33]	0.435	0.208	0.192	0.051	0.077
DR-AVIT [7]	0.206	0.088	0.054	0.012	0.021
PID [22]	0.673	0.577	0.601	0.197	0.269
DiffV2IR [23]	0.765	0.694	0.678	0.158	0.272
ThermalGen [34]	0.855	0.681	0.766	0.315	0.372
CoVisIT (Ours)	**0.970**	**0.941**	**0.981**	**0.792**	**0.676**
HR GT	0.976	0.957	0.987	0.838	0.712

Computational efficiency. Table 4 compares model parameters, FLOPs, and inference time for 512×512 inputs on a single NVIDIA A100 GPU. The Transformer-based method (SwinFuSR) and GAN-based methods (DR-AVIT and StegoGAN) have fewer parameters and lower FLOPs than the autoregressive method (Real-IISR), flow-based method (ThermalGen), and diffusion-based methods (3M-TI, PID, DiffV2IR, and CoVisIT). However, parameter counts and FLOPs do not directly determine inference time. For example, 3M-TI and Real-IISR process each image in less than 0.5 s, whereas CoVisIT and DiffV2IR require 4.550 s and 5.183 s, with 50-step diffusion sampling.

Although CoVisIT does not offer a computational efficiency advantage, its iterative denoising process progressively integrates complementary cross-modal priors to generate IR images with faithful thermal distributions and accurate spatial alignment with HR VIS images. Efficiency improvements could be achieved by reducing sampling steps and simplifying computations within each step. To this end, future work will explore consistency-model distillation to shorten the sampling trajectory, caching of reusable latent features to avoid redundant computation, and lightweight denoising networks to reduce per-step costs while preserving generation quality.

Table 4: Computational efficiency comparison.

Method	Params (M) ↓	FLOPs (G) ↓	Time (s) ↓
Real-IISR [42]	1253.61	15842.5	0.400
SwinFuSR [1]	**1.71**	1065.4	2.360
3M-TI [2]	1307.19	9089.4	0.261
StegoGAN [33]	12.56	471.4	0.021
DR-AVIT [7]	26.16	**109.1**	**0.011**
PID [22]	309.21	13857.0	0.864
DiffV2IR [23]	1066.26	118852.0	5.183
ThermalGen [34]	625.15	61565.9	1.092
CoVisIT (Ours)	824.42	136700.1	4.550

Ablation study. To verify the effectiveness of each specific design, we conduct ablation studies on LLVIP, with results reported in Tab. 5. Compared with the baseline using only VIS input, introducing LR IR thermal-distribution cues brings clear improvements, confirming the importance of explicit thermal references. Building on this, CMA further improves performance by reducing modality representation gaps for more reliable cross-modal prior extraction. MSDCM further enhances most metrics, demonstrating the benefit of input-adaptive local alignment under VIS-IR misalignment. The full model with text guidance achieves the best overall performance, indicating that auxiliary scene descriptions further help thermal target understanding. Qualitative ablation results and detailed visual analyses are provided in the Appendix.

Table 5: Quantitative results of ablation study.

Config	VIS	IR	CMA	MSDCM	Text	PSNR↑	SSIM↑	FSIM↑	VIF↑	LPIPS↓	FID↓
I	✓					21.546	0.810	0.904	0.248	0.178	66.463
II	✓	✓				26.968	0.856	0.926	0.277	0.122	44.181
III	✓	✓	✓			27.713	0.864	0.932	0.289	0.124	46.088
IV	✓	✓	✓	✓		29.057	0.891	0.948	0.340	0.118	46.550
Ours	✓	✓	✓	✓	✓	**29.872**	**0.906**	**0.956**	**0.372**	**0.110**	**43.437**

Limitations and discussions. Although CoVisIT achieves promising results, a gap remains between simulated and real degradations. The misaligned LR IR inputs are mainly constructed by downsampling, affine perturbation, and elastic deformation, whereas real IR imaging may involve sensor noise, nonuniform responses, and complex viewpoint changes. These factors may weaken thermal-distribution cues and make local alignment more challenging, motivating future work on realistic degradation modeling. Beyond these limitations, CoVisIT explores the integration of complementary VIS and IR information with HR IR images as the output. This formulation could be extended to reconstruct high-quality VIS counterparts under degraded conditions using complementary IR cues. A further direction is to jointly generate high-quality, spatially aligned VIS and IR images.

5 Conclusion

In this paper, we propose **CoVisIT**, a diffusion-based framework with cross-modal prior guidance for visible-to-infrared translation. CoVisIT bridges visible texture generation and infrared thermal fidelity by jointly leveraging HR VIS images and misaligned LR IR inputs within a unified diffusion framework. Specifically, HR VIS defines the output field of view and geometry, while LR IR provides thermal-distribution cues. To effectively exploit complementary cross-modal priors, CMA reduces modality representation gaps and enables reliable prior interaction, while DKGM and MSDCM perform input-adaptive local alignment to handle spatial misalignment during denoising. Extensive experiments demonstrate that CoVisIT generates VIS-aligned, HR IR images with clearer structures and more faithful thermal distributions. The generated images also benefit downstream vision tasks, including image fusion, semantic segmentation, and object detection.

Acknowledgments and Disclosure of Funding

This work was supported by NSFC (62602480, 62606372, 62676289) and CPSF (GZB20260061, 2026M791669). The numerical calculations have been done on the supercomputing system in the Supercomputing Center of Wuhan University. The authors declare no competing interests.

References

[1] Cyprien Arnold, Philippe Jouvet, and Lama Seoud. SwinFuSR: An image fusion-inspired model for RGB-guided thermal image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 3027–3036, 2024.

[2] Minchong Chen, Xiaoyun Yuan, Junzhe Wan, Jianing Zhang, and Jun Zhang. 3M-TI: High-quality mobile thermal imaging via calibration-free multi-camera cross-modal diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 5659–5669, 2026.

[3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 184–199, 2014.

[4] Wenhao Dong, Haodong Zhu, Shaohui Lin, Xiaoyan Luo, Yunhang Shen, Guodong Guo, and Baochang Zhang. Fusion-mamba for cross-modality object detection. *IEEE Transactions on Multimedia*, 27: 7392–7406, 2025.

[5] Meiqi Gong, Hao Zhang, Bingwei Hui, and Jiayi Ma. TransFM: Visible-to-infrared image translation via flow matching. *IEEE/CAA Journal of Automatica Sinica*, 13(5):1239–1241, 2026.

[6] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-min Hu. SegNeXt: Rethinking convolutional attention design for semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 1140–1156, 2022.

[7] Zonghao Han, Shun Zhang, Yuru Su, Xiaoning Chen, and Shaohui Mei. DR-AVIT: Toward diverse and realistic aerial visible-to-infrared image translation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

[8] 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*, pages 6626–6637, 2017.

[9] Deepak Kumar Jain, Xudong Zhao, Germán González-Almagro, Chenquan Gan, and Ketan Kotecha. Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes. *Information Fusion*, 95:401–414, 2023.

[10] 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*, pages 3496–3504, 2021.

[11] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Yolov8. `https://github.com/ultralytics/ultralytics`, 2023. Ultralytics, version 8.x.

[12] Aditya Kasliwal, Pratinav Seth, Sriya Rallabandi, and Sanchit Singhal. CoReFusion: Contrastive regularized fusion for guided thermal super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 507–514, 2023.

[13] Vladimir V. Kniaz, Vladimir A. Knyaz, Jiri Hladuvka, Walter G. Kropatsch, and Vladimir Mizginov. ThermalGAN: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 606–624, 2018.

[14] 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 *Proceedings of the European Conference on Computer Vision*, pages 236–256, 2024.

[15] Dong-Guw Lee, Myung-Hwan Jeon, Younggun Cho, and Ayoung Kim. Edge-guided multi-domain RGB-to-TIR image translation for training vision tasks with challenging labels. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 8291–8298, 2023.

[16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7534–7544, 2025.

[17] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1833–1844, 2021.

[18] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022.

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

[20] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, Xin Fan, and Risheng Liu. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2025.

[21] Ying Lv, Zhi Liu, and Gongyang Li. Context-aware interaction network for RGB-T semantic segmentation. *IEEE Transactions on Multimedia*, 26:6348–6360, 2024.

[22] Fangyuan Mao, Jilin Mei, Shun Lu, Fuyang Liu, Liang Chen, Fangzhou Zhao, and Yu Hu. PID: Physics-informed diffusion model for infrared image generation. *Pattern Recognition*, 169:111816, 2026.

[23] Lingyan Ran, Lidong Wang, Guangcong Wang, Peng Wang, and Yanning Zhang. DiffV2IR: visible-to-infrared diffusion model via vision-language understanding. *arXiv preprint arXiv:2503.19012*, 2025.

[24] Rafael E Rivadeneira, Patricia L Suárez, Angel D Sappa, and Boris X Vintimilla. Thermal image superresolution through deep convolutional neural network. In *Proceedings of the International Conference on Image Analysis and Recognition*, pages 417–426, 2019.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[26] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.

[27] Linrui Shi, Gaochang Wu, Yingqian Wang, Yebin Liu, and Tianyou Chai. DuaDiff: Dual-conditional diffusion model for guided thermal image super-resolution. *IEEE Transactions on Neural Networks and Learning Systems*, 37(6):2809–2823, 2026.

[28] 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.

[29] Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Deep learning-based image fusion: A survey. *Journal of Image and Graphics*, 28(1):3–36, 2023.

[30] Emma Wadsworth, Advait Mahajan, Raksha Prasad, and Rajesh Menon. Deep learning for thermal-RGB image-to-image translation. *Infrared Physics & Technology*, 141:105442, 2024.

[31] Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009.

[32] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[33] Sidi Wu, Yizi Chen, Samuel Mermet, Lorenz Hurni, Konrad Schindler, Nicolas Gonthier, and Loic Landrieu. StegoGAN: Leveraging steganography for non-bijective image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7922–7931, 2024.

[34] Jiuhong Xiao, Roshan Nayak, Ning Zhang, Daniel Tortei, and Giuseppe Loianno. ThermalGen: Style-disentangled flow-based generative models for RGB-to-thermal image translation. In *Advances in Neural Information Processing Systems*, pages 33111–33133, 2025.

[35] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of clip. In *Proceedings of the European Conference on Computer Vision*, pages 310–325, 2024.

[36] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24(12):14679–14694, 2023.

[37] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.

[38] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[39] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25912–25921, 2024.

[40] Hang Zhong, Yu Wang, and Shengjie Zhao. SwinPaste: A swin transformer-based framework for RGB-guided thermal image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4628–4633, 2025.

[41] Yan Zou, Linfei Zhang, Chengqian Liu, Bowen Wang, Yan Hu, and Qian Chen. Super-resolution reconstruction of infrared images based on a convolutional neural network with skip connections. *Optics and Lasers in Engineering*, 146:106717, 2021.

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16365–16375, 2026.

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

A Appendix

A.1 Gemini prompt template for scene description generation

This appendix presents the prompt used with Gemini to generate structured scene descriptions from LR IR images. Specifically, Gemini is employed to analyze the thermal infrared images and extract structured textual descriptions characterizing the scene context, visible high-temperature objects, their spatial distributions, and inter-object relationships. To ensure consistency and reduce ambiguity, the prompt explicitly constrains the output vocabulary, sentence structures, and spatial relation expressions. The resulting descriptions T_{ir} are subsequently encoded by Long-CLIP to obtain the text condition $\mathbf{c}_{ir}$, which is incorporated into the conditional denoising process. The complete prompt template is presented below.

Gemini Prompt

Analyze the thermal infrared image. First, briefly describe the scene context using a short scene word or phrase, such as street, square, woodland, parking lot, road, field, building area, waterside, or indoor. Use the sentence pattern: "The scene is a/an + scene description."

Then, describe only the visible high-temperature objects. If the object type is uncertain, classify it as a heat source. Use only number words from one to twenty for counts, rather than words such as multiple, several, or many. Object descriptions must strictly follow three sentence types.

1) Quantity sentence: *number word + object category + is/are present.* Use the singular category with "is" for one and the plural category with "are" for two to twenty, e.g., "One human is present" or "Two vehicles are present."

2) Object-distribution sentence: *The + singular/plural category + is/are + state relation.* "Clustered" indicates that multiple objects of the same category are grouped; "separated" indicates that objects are spatially dispersed; "isolated" indicates that one object is clearly separated from other high-temperature objects; and "together" indicates that multiple objects occupy a nearby region, e.g., "The humans are clustered" or "The vehicle is isolated."

3) Inter-object relation sentence: *The + category 1 + is/are + spatial relation + the + category 2.* "Near" indicates spatial proximity; "around" indicates that one or more objects surround another object; and "adjacent to" indicates spatial adjacency, e.g., "The human is near the vehicle" or "The humans are around the heat source."

Each sentence should describe only one quantity, distribution state, or inter-object relation. Different object categories should be described in separate sentences. Quantity and relation information should not be combined within the same sentence. Do not introduce other sentence structures or relation expressions, add adjectives, modify category names, or describe actions, poses, sizes, shapes, or attributes. Except for the initial scene sentence, ignore the background and low-temperature regions. Avoid repeated information.

Output a single line, with sentences separated by periods and no line breaks.

A.2 Misalignment simulation details

We provide additional details of the misalignment simulation in this section. This simulation mimics the spatial mismatch between VIS and IR images caused by different sensor positions, viewpoint changes, and local parallax. In our implementation, the IR input is first degraded by downsampling and then warped by a combination of global affine perturbation and local elastic deformation.

Let I_{ir}^{u} denote the upsampled LR IR image to be warped, and let $\mathbf{G} \in [-1,1]^{\bar{H}\times\bar{W}\times 2}$ be its normalized coordinate grid. For each spatial position $\mathbf{x}$, the misalignment process first constructs a global affine perturbation. Following the implementation, the rotation range is normalized according to the image size and the base scale b:

$$\mathbf{s}_b = \left(\frac{b}{\bar{W}}, \frac{b}{\bar{H}}\right), \qquad \bar{\theta} = \frac{\max(\bar{H},\bar{W})}{b}\theta, \tag{14}$$

where θ denotes the maximum rotation angle, and b is the base value used to normalize the deformation magnitude across different resolutions. The random rotation angle and translation vector are sampled as:

$$\theta_r \sim \mathcal{U}(-\bar{\theta},\bar{\theta}), \qquad \mathbf{t}_r = (t_x, t_y), \quad t_x, t_y \sim \mathcal{U}(-\tau,\tau), \tag{15}$$

where τ denotes the maximum translation range. The affine-warped coordinate grid is then computed by the inverse affine transformation:

$$\mathbf{G}_{aff}(\mathbf{x}) = \mathbf{A}_{\theta_r,\mathbf{t}_r}^{-1}\mathbf{G}(\mathbf{x}), \tag{16}$$

where $\mathbf{A}_{\theta_r,\mathbf{t}_r}$ is the affine matrix parameterized by θ_r and $\mathbf{t}_r$. The inverse form is used because image warping samples source coordinates for each target position.

In addition to the global affine perturbation, we further generate a local elastic displacement field. A two-channel random displacement noise field is first sampled from a uniform distribution:

$$\mathbf{N}(\mathbf{x}) = [N_x(\mathbf{x}), N_y(\mathbf{x})], \qquad N_x(\mathbf{x}), N_y(\mathbf{x}) \sim \mathcal{U}(-1, 1), \tag{17}$$

and then smoothed by Gaussian filtering to obtain a spatially continuous noise field:

$$\bar{\mathbf{N}} = (\mathcal{B}_\sigma)^6(\mathbf{N}), \tag{18}$$

where $(\mathcal{B}_\sigma)^6$ denotes six consecutive Gaussian smoothing operations with standard deviation σ. The elastic displacement field is then obtained as:

$$\mathbf{D}_{ela}(\mathbf{x}; r, \sigma) = r \cdot \bar{\mathbf{N}}(\mathbf{x}), \tag{19}$$

where r corresponds to the `ratio` parameter and controls the strength of the elastic deformation. In the implementation, the Gaussian kernel size is determined by:

$$k_\sigma = 2\left\lfloor \frac{3\sigma}{2} \right\rfloor + 1. \tag{20}$$

By combining the affine perturbation and elastic displacement field, the final displacement field is formulated as:

$$\Delta\mathbf{G}(\mathbf{x}; \theta, \tau, r, \sigma, b) = \mathbf{s}_b \odot \left[\mathbf{D}_{ela}(\mathbf{x}; r, \sigma) + \mathbf{G}_{aff}(\mathbf{x}; \theta, \tau, b) - \mathbf{G}(\mathbf{x})\right], \tag{21}$$

where $\odot$ denotes element-wise multiplication. The final transformed sampling grid is:

$$\mathbf{G}'(\mathbf{x}) = \mathbf{G}(\mathbf{x}) + \Delta\mathbf{G}(\mathbf{x}; \theta, \tau, r, \sigma, b). \tag{22}$$

The warped IR input can then be written as:

$$\widetilde{I}_{ir}^{u}(\mathbf{x}) = I_{ir}^{u}\left(\mathbf{G}'(\mathbf{x})\right). \tag{23}$$

When $\mathbf{G}'(\mathbf{x})$ falls outside the valid normalized coordinate range $[-1, 1]$, the corresponding position is regarded as an out-of-boundary region and marked by a binary mask:

$$\mathbf{M}_{out}(\mathbf{x}) = \mathbb{I}\left[\mathbf{G}'_x(\mathbf{x}) < -1 \ \vee \ \mathbf{G}'_x(\mathbf{x}) > 1 \ \vee \ \mathbf{G}'_y(\mathbf{x}) < -1 \ \vee \ \mathbf{G}'_y(\mathbf{x}) > 1\right]. \tag{24}$$

In the qualitative deformation experiment shown in Fig. 4, we define three misalignment levels by fixing the global affine perturbation and Gaussian smoothing scale, while only changing the elastic deformation strength. Specifically, the shared parameters are set as:

$$\theta = 12^\circ, \qquad \tau = 0.06, \qquad \sigma = 15, \qquad b = 500. \tag{25}$$

The elastic deformation strength r is configured as:

$$\begin{aligned} &\text{Level 1}: \quad r = 0.5, \\ &\text{Level 2}: \quad r = 1.0, \\ &\text{Level 3}: \quad r = 1.8. \end{aligned} \tag{26}$$

Thus, the three levels mainly differ in local non-rigid deformation strength. As r increases, the local displacement field is amplified, leading to stronger bending and spatial misalignment around object boundaries, building contours, and background structures. This setting provides progressively more challenging scenarios for evaluating the robustness of VIS-to-IR generation methods.

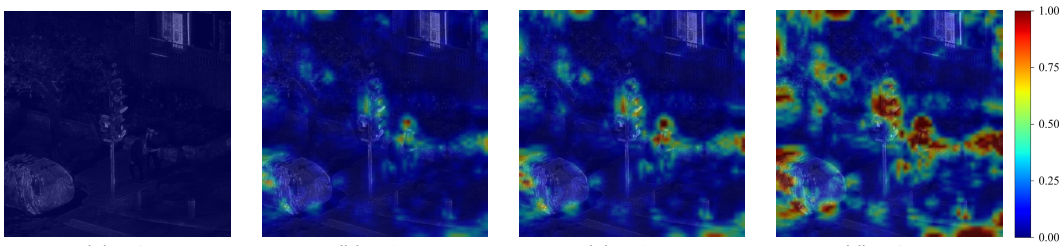


Figure 6: Visualization of IR input feature variations before the 5×5 dynamic convolution branch at Level 4 under different elastic deformation strengths ($r = 0, 1, 1.5, 3$).

A.3 Visualization of IR feature variations under elastic deformation

To investigate the impact of spatial misalignment on the features processed by MSDCM, we visualize the variations in IR input features under different elastic deformation strengths. Specifically, we keep the HR VIS input and the underlying elastic displacement field fixed, while varying only the elastic deformation strength $r \in \{0, 1, 1.5, 3\}$, where $r = 0$ denotes the undeformed reference. We select the second MSDCM at Level 4 of the denoising U-Net and extract its IR input features before the 5×5 dynamic convolution branch. For each deformation strength, we compute the feature variations relative to the corresponding features at $r = 0$ and visualize the resulting spatial differences on the same VIS reference image using a unified color scale.

As shown in Fig. 6, the feature variations are zero at $r = 0$, since this setting serves as the reference. The color bar indicates the normalized magnitude of feature variation, ranging from blue (low) to red (high). As the deformation strength increases, more pronounced feature variations emerge around local scene structures, accompanied by an expansion of high-response regions. Under severe deformation ($r = 3$), significant feature variations extend over broader spatial regions. These observations indicate that increasing spatial deformation substantially affects both the magnitude and spatial distribution of the IR features entering MSDCM, further motivating its input-adaptive dynamic convolution design for local feature aggregation.

A.4 Robustness to diverse misalignments

To further evaluate robustness to diverse VIS–IR misalignment conditions, we conduct controlled experiments by independently varying the **rotation angle**, **translation ratio**, and **elastic deformation ratio**. Their default training settings are 5°, 0.05, and 0.5, respectively. We vary rotation from 0° to 10°, translation from 0 to 0.10, and elastic deformation from 0 to 1.20, while fixing the other two parameters at their default settings. We evaluate CoVisIT and 3M-TI using the same misalignment settings, yielding 15 parameter-sweep entries.

As shown in Table 6, both methods exhibit performance degradation as misalignment severity increases. Nevertheless, CoVisIT consistently outperforms 3M-TI across all six metrics in every tested setting. Even when rotation, translation, or elastic deformation exceeds the training range, CoVisIT maintains more faithful IR generation. Across the 15 parameter-sweep entries, CoVisIT achieves average PSNR and SSIM values of 29.434 and 0.897, compared with 24.567 and 0.760 for 3M-TI. These results demonstrate that CoVisIT remains robust across diverse spatial misalignments, including those beyond the training range.

A.5 Qualitative and quantitative results of multi-modal vision task

Image fusion. As shown in Fig. 5, different IR generation results directly affect target saliency, structural clarity, and thermal-response consistency in the fused images. For the IR SR method, as it only relies on the LR IR input, it is difficult to achieve pixel-wise alignment with the VIS image. As a result, obvious artifacts can be observed in the green box of the first row in Fig. 5 (b). Although VIS-guided IR SR methods can introduce structural information from the VIS modality, they do not sufficiently consider the representation gap between VIS and IR modalities and local spatial misalignment. Therefore, their fusion results still suffer from blurred scene structures and inconsistent thermal responses, as shown in Fig. 5 (c). For VIS-to-IR translation methods, the generated IR images

Table 6: Robustness comparison under varying misalignment conditions. When one parameter varies, the other two are fixed at their default training settings. The best average results are shown in bold.

Parameter	Level	PSNR↑		SSIM↑		FSIM↑		VIF↑		LPIPS↓		FID↓	
		CoVisIT	3M-TI	CoVisIT	3M-TI	CoVisIT	3M-TI	CoVisIT	3M-TI	CoVisIT	3M-TI	CoVisIT	3M-TI
Rotation	0°	30.182	25.756	0.911	0.785	0.960	0.872	0.382	0.162	0.107	0.338	42.472	116.135
	2.5°	30.071	25.648	0.910	0.784	0.959	0.870	0.377	0.161	0.108	0.342	42.931	119.836
	5°	29.872	24.707	0.906	0.765	0.956	0.854	0.372	0.142	0.110	0.363	43.437	124.324
	7°	29.215	24.124	0.894	0.749	0.947	0.837	0.351	0.127	0.117	0.373	44.927	130.784
	10°	27.362	22.631	0.845	0.710	0.914	0.805	0.294	0.100	0.140	0.401	47.782	137.989
Translation	0	30.246	25.920	0.912	0.797	0.961	0.879	0.383	0.180	0.107	0.329	42.388	120.465
	0.025	30.119	25.631	0.911	0.790	0.960	0.873	0.380	0.169	0.109	0.338	42.744	121.049
	0.05	29.872	24.707	0.906	0.765	0.956	0.854	0.372	0.142	0.110	0.363	43.437	124.324
	0.075	29.288	23.790	0.898	0.737	0.950	0.828	0.357	0.111	0.116	0.386	44.177	131.631
	0.10	27.545	22.379	0.856	0.698	0.923	0.794	0.297	0.081	0.131	0.420	46.392	141.286
Elastic	0	30.431	25.570	0.914	0.785	0.962	0.868	0.395	0.163	0.107	0.354	42.681	121.084
	0.25	30.269	25.432	0.913	0.782	0.960	0.865	0.387	0.159	0.109	0.355	42.973	123.118
	0.50	29.872	24.707	0.906	0.765	0.956	0.854	0.372	0.142	0.110	0.363	43.437	124.324
	0.80	29.146	24.298	0.897	0.757	0.949	0.845	0.347	0.128	0.115	0.365	44.249	127.554
	1.20	28.027	23.204	0.879	0.729	0.937	0.824	0.309	0.104	0.119	0.366	45.908	133.019
Average		**29.434**	24.567	**0.897**	0.760	**0.950**	0.848	**0.358**	0.138	**0.114**	0.364	**43.996**	126.461

Table 7: Quantitative comparison of image fusion results on LLVIP, FMB, M3FD, and MSRS using the pretrained EMMA model.

Method	EN↑	SF↑	SD↑	MI↑	VIF↑	Q_{abf}↑	PSNR↑	SSIM↑	LPIPS↓
Real-IISR [42]	6.988	17.789	41.782	3.377	0.642	0.530	31.592	0.936	0.060
SwinFuSR [1]	6.974	17.563	41.673	3.240	0.632	0.525	26.029	0.880	0.121
3M-TI [2]	6.996	17.678	41.929	3.401	0.645	0.527	30.558	0.925	0.077
PID [22]	6.989	**18.245**	41.330	3.181	0.596	0.510	26.128	0.899	0.068
DiffV2IR [23]	7.004	18.014	41.871	3.153	0.575	0.504	25.907	0.873	0.080
ThermalGen [34]	6.997	18.158	41.710	3.275	0.625	0.523	27.866	0.924	0.057
CoVisIT (Ours)	**7.008**	18.078	**42.121**	**3.494**	**0.669**	**0.541**	**35.346**	**0.967**	**0.031**
HR GT	7.012	18.586	42.074	3.638	0.731	0.595	–	–	–

tend to contain ill-posed thermal distributions because they mainly infer IR appearance from VIS-centric priors. This leads to inaccurate thermal-response distributions in Figs. 5 (e), (f), and (g), which cannot accurately correspond to object structures in the VIS images. In contrast, our method jointly exploits scene-detail priors from the VIS image and thermal-distribution priors from the LR IR input, while alleviating modality gaps and local misalignment through CMA and MSDCM. Therefore, the fusion result in Fig. 5 (h) shows better target saliency, clearer edges, and more consistent thermal responses, making it closer to the HR GT fusion result in Fig. 5 (i).

To quantitatively evaluate the downstream fusion performance, we conduct experiments on 2,246 image pairs from LLVIP, FMB, M3FD, and MSRS using the same pretrained EMMA [39] model. We adopt nine evaluation metrics, including EN, SF, and SD as no-reference metrics, MI, VIF, and Q_{abf} for source information preservation, and *PSNR, SSIM, and LPIPS for comparison with the reference fusion obtained using HR GT IR and HR VIS images*. As shown in Tab. 7, CoVisIT achieves the best performance on eight of the nine metrics among all compared generation methods. In particular, its superior PSNR, SSIM, and LPIPS scores demonstrate that the fusion results based on CoVisIT-generated IR images are more consistent with the reference fusion results.

Semantic segmentation. High-quality IR images are also important for semantic segmentation, since pixel-level prediction requires generated images to provide clear object boundaries, stable regional responses, and structural information consistent with real object contours. We use a SegNeXt segmentation model trained on real HR IR images to compare segmentation results obtained from different generated IR images. As shown in Fig. 7, the IR SR method Real-IISR produces relatively continuous segmentation regions. However, due to the lack of structural textures, its segmentation

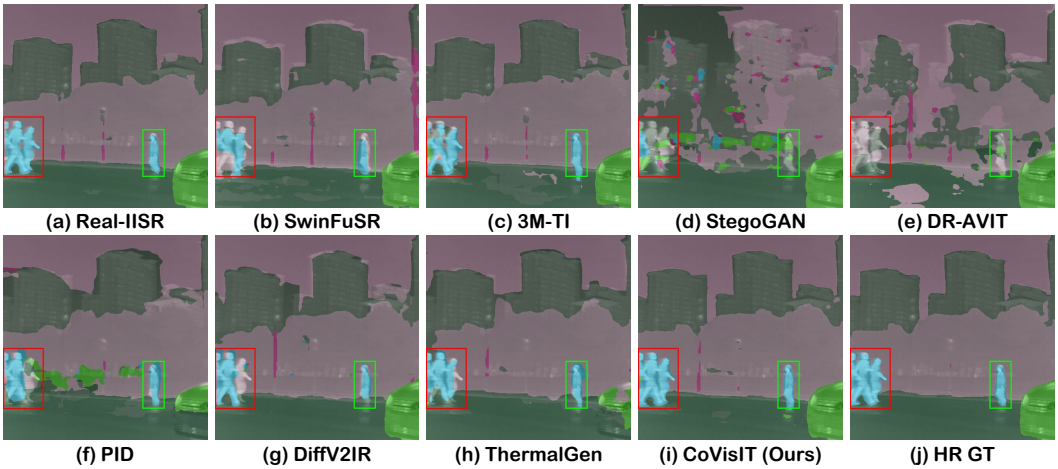


Figure 7: Qualitative semantic segmentation results on FMB.

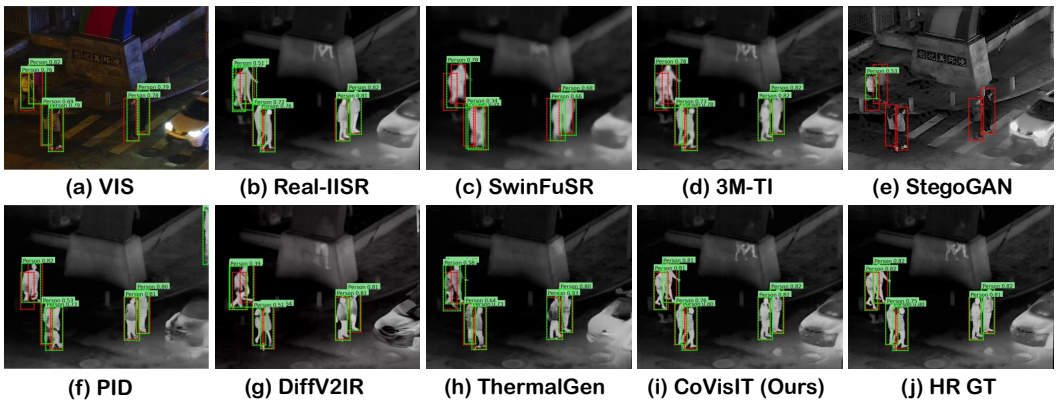


Figure 8: Qualitative object detection results on LLVIP.

result does not fully match real pedestrian boundaries and fails to accurately cover target contours, as highlighted in Fig. 7 (a). After introducing VIS guidance, SwinFuSR and 3M-TI can supplement partial scene structural information, but the object boundaries and thermal responses in their generated results are still not sufficiently reliable. As a result, the segmentation maps only cover part of the real pedestrian regions, as shown in Fig. 7 (b) and Fig. 7 (c). For VIS-to-IR translation methods, since they mainly infer IR information from VIS appearance, the thermal semantics in the generated results tend to be unstable, which further affects pixel-level prediction. Specifically, Figs. 7 (d), (e), (f), and (h) show discontinuous segmented regions, insufficient target coverage, category misclassification, and unstable correspondence between thermal responses and real target regions, respectively. In contrast, CoVisIT jointly exploits scene-detail priors from the VIS image and thermal-distribution priors from the LR IR input, producing IR images with clearer object boundaries and more reliable thermal-semantic consistency. Therefore, the segmentation result in Fig. 7 (i) is closer to the result of the real HR IR image in Fig. 7 (j), especially in terms of region continuity, boundary completeness, and pedestrian coverage.

Object detection. Object detection further examines whether generated IR images preserve discriminative thermal cues and localization-friendly object structures. We use a YOLOv8 detector trained on real HR IR images to evaluate different generated IR images, with red boxes indicating ground-truth annotations and green boxes denoting YOLOv8 predictions, thereby enabling a visual comparison of detection performance across different methods. As shown in Fig. 8 (b), although Real-IISR restores the image from the LR IR input and preserves certain thermal information, its local object boundaries remain blurred and redundant responses may be introduced, leading to an additional false detection in the left pedestrian group. SwinFuSR and 3M-TI introduce the VIS image as structural guidance, but

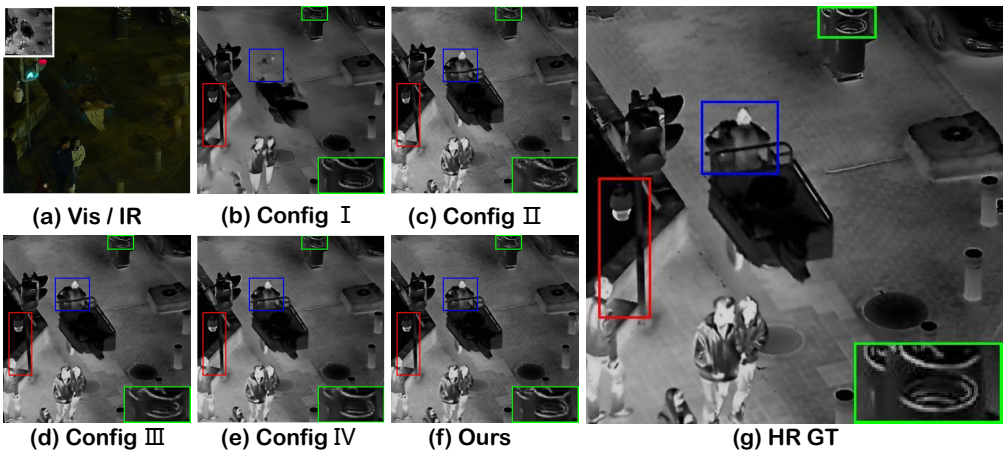


Figure 9: Qualitative ablation results of different configurations.

the cross-modal information is still not fully exploited, and the local spatial inconsistency between IR and VIS images is not completely resolved. Therefore, missed detections can be observed in Figs. 8 (c) and (d), where the predicted boxes also show clear deviations in position and scale from the real targets. For VIS-to-IR generative methods, *i.e.,* StegoGAN, PID, DiffV2IR, and ThermalGen, the IR thermal distribution is mainly estimated from the VIS appearance. The unstable thermal responses of the generated targets reduce the detection confidence in Figs. 8 (e)–(h). In contrast, by fully exploiting complementary scene priors, CoVisIT generates IR images with more stable target responses and clearer pedestrian contours, making the targets easier to localize and distinguish. As a result, the detection result in Fig. 8 (i) is more stable in terms of target number, box location, and confidence, and is closer to the detection result of the real HR IR image in Fig. 8 (j).

A.6 Qualitative results of ablation study

Figure 9 provides qualitative ablation results for different configurations. In Config **I**, where only the VIS image is used as input, the generated result suffers from incorrect thermal estimation in the blue box. This is because, under low-light conditions, the VIS image provides weak evidence for recognizing the person in this region, making the corresponding thermal response hard to infer. After introducing the misaligned LR IR input in Config **II**, the thermal-distribution error is largely reduced. However, due to the spatial misalignment of the LR IR cues, clear distortion can still be observed around the pole in the left red box. Meanwhile, the large modality gap also introduces noisy edges around the trash bin in the upper green box. With CMA in Config **III**, the modality representation gap is reduced and complementary cross-modal priors are better exploited. As a result, the edge structures become clearer, and the distortion in the left red box is also alleviated. After further introducing MSDCM in Config **IV**, the local distortion is significantly reduced, showing the effectiveness of dynamic local alignment for handling spatial misalignment. Finally, with text guidance, the full model achieves the best visual quality, producing more faithful thermal responses and clearer local structures. These results verify the complementary contributions of LR IR cues, CMA, MSDCM, and text guidance to faithful thermal response recovery and structure-aware IR generation.

A.7 Broader impacts

CoVisIT aims to generate HR IR images aligned with VIS images, which can benefit robust multi-modal perception in applications such as autonomous driving, robotics, intelligent monitoring, and nighttime scene understanding. By recovering clearer structures and more reliable thermal responses, the generated IR images can provide useful inputs for downstream tasks, including image fusion, semantic segmentation, and object detection. This may help improve visual perception under challenging conditions such as low illumination, adverse weather, and complex urban scenes. This work is evaluated on public research benchmarks and does not involve private or sensitive personal data. Potential real-world deployment in monitoring and safety-critical scenarios should comply with appropriate privacy, safety, and ethical regulations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the main contribution and scope of the work: a cross-modal prior guided diffusion framework that jointly exploits visible texture priors and low-resolution infrared thermal-distribution cues to generate visible-aligned high-resolution infrared images with clearer structures and more reliable thermal responses. The claims are supported by quantitative and qualitative comparisons, ablation studies, robustness evaluation, and downstream vision task results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a limitation discussion, pointing out the limits of CoVisIT in realistic degradation modeling and inference efficiency. Specifically, the misaligned low-resolution infrared inputs are mainly constructed by simulated degradations, while real infrared imaging may involve more complex sensor noise, nonuniform responses, and viewpoint changes. Diffusion-based generation also remains slower than feed-forward models.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not present theoretical results, theorems, or formal proofs. The proposed method is an empirical framework whose effectiveness is validated through quantitative comparisons, qualitative results, and ablation studies.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper reports the key information needed to reproduce the main results, including the network architecture, training datasets, data simulation details, preprocessing, evaluation metrics, compared methods, training hyperparameters, sampling settings, and implementation environment. In addition, the source code and pretrained models are publicly available at `https://github.com/Linfeng-Tang/CoVisIT` to further support reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The source code and the pretrained models are publicly available at `https://github.com/Linfeng-Tang/CoVisIT` to support reproducibility. The paper describes the public datasets, preprocessing and misalignment simulation details, evaluation metrics, compared methods, training settings, sampling settings, and implementation environment needed to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the training and test details needed to understand the results, including the training datasets and test sets, mixed training setup, preprocessing and data augmentation, misalignment simulation settings, optimizer, learning rate, diffusion training steps, DDIM sampling steps, evaluation metrics, and compared methods. The architecture and hardware environment are also reported, and additional implementation details are provided in the appendix.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This paper does not include error bars or formal statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the compute resources and execution time required for the experiments, including the CPU, GPU type, GPU memory, implementation platform, and running time for the proposed method. These details provide sufficient information to assess the computational cost and reproduce the reported experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms to the NeurIPS Code of Ethics. It uses public visible-infrared image datasets for research on infrared image generation and downstream multi-modal perception, and does not involve human-subject experiments, private or sensitive personal data, or data collection from users.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses the broader impacts of the proposed visible-to-infrared image generation method, mainly including its potential benefits for robust multi-modal perception in autonomous driving, intelligent surveillance, robotics, and nighttime scene understanding. Since this work focuses on high-quality infrared image generation and downstream vision tasks, no direct negative societal impacts are identified.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work does not release high-misuse-risk datasets or general-purpose generative models. Although CoVisIT is a generative framework, it is designed for visible-to-infrared image translation on public research benchmarks, rather than general image generation or scraped datasets. Therefore, additional safeguards are not required.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite the original papers of the datasets and baseline methods used in this work. All public datasets and existing assets are used only for academic research and are handled in accordance with their original licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces new code and model assets, including the implementation code and pretrained models of CoVisIT. Both the implementation code and the pretrained models are publicly available at `https://github.com/Linfeng-Tang/CoVisIT`, with documentation and instructions for usage, reproduction, and evaluation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing experiments, human-subject studies, participant instructions, or compensation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing experiments, human-subject studies, study participants, or personally identifiable data. Therefore, IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper discloses the use of Gemini API as an auxiliary vision-language tool for extracting scene descriptions from low-resolution infrared inputs. This usage is limited to generating auxiliary text conditions and is not used to develop the core idea, design the proposed modules, conduct experimental analysis, or formulate scientific claims. Any additional LLM usage is limited to language polishing, editing, and formatting assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.