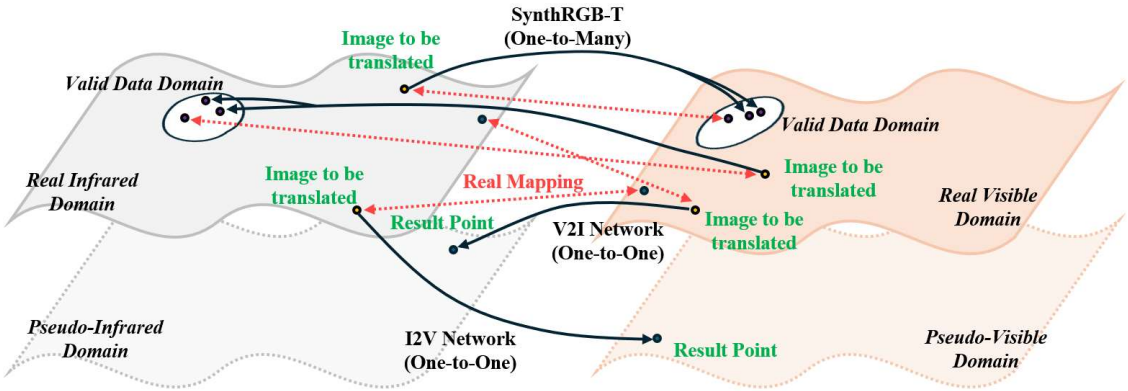

# SynthRGB-T: Language-Vision Guided Image Translation for Diversity Synthesis

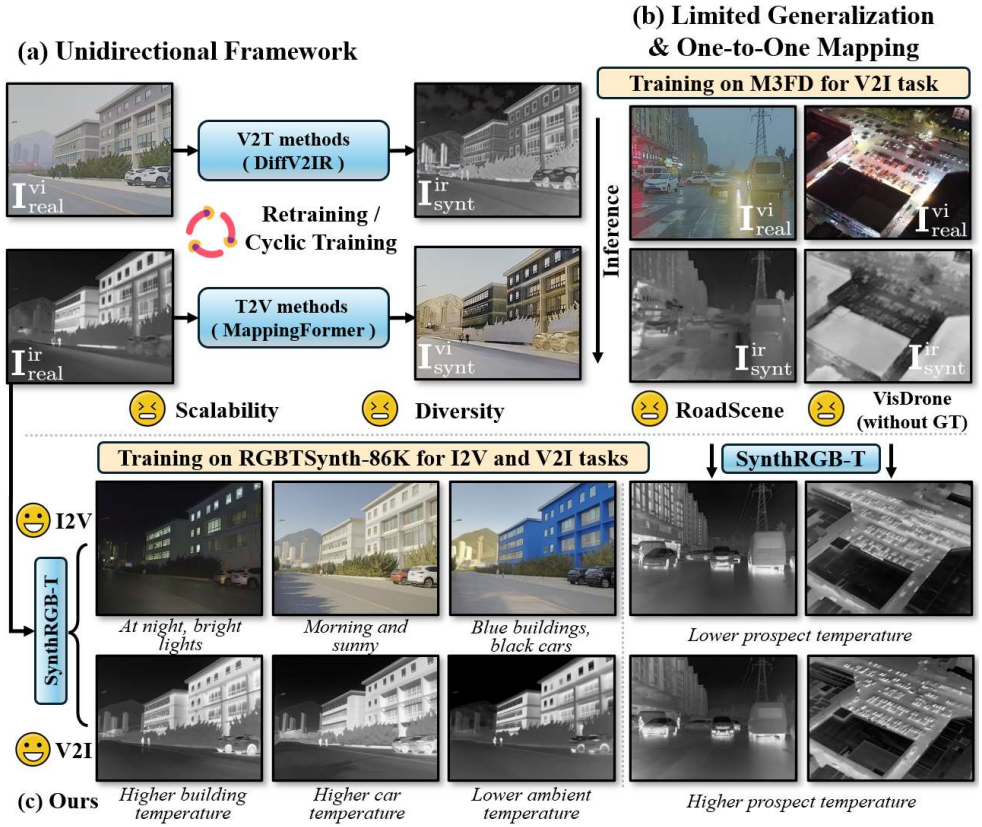
Jiangang Ding[1] Yiquan Du[1] Pengxiang Li[2] Lili Pei[1] Yuanlin Zhao[1,†] Wei Li[1,†]

[1]Chang'an University [2]Hong Kong Polytechnic University


# [Motivation]

- **One-way modality mapping**: Most methods rely on fixed frameworks for one-way mapping, and they need repeated retraining to meet various generation requirements.
- **Limited scalability**: Owing to the absence of explicit modeling for open scenarios, these methods struggle to capture the correspondence between visible and thermal signals, often converging to suboptimal solutions constrained by the training benchmarks.
- **Lack of diversity**: Existing methods mainly learn one-to-one mappings, which limits their ability to capture the inherent diversity of possible samples.

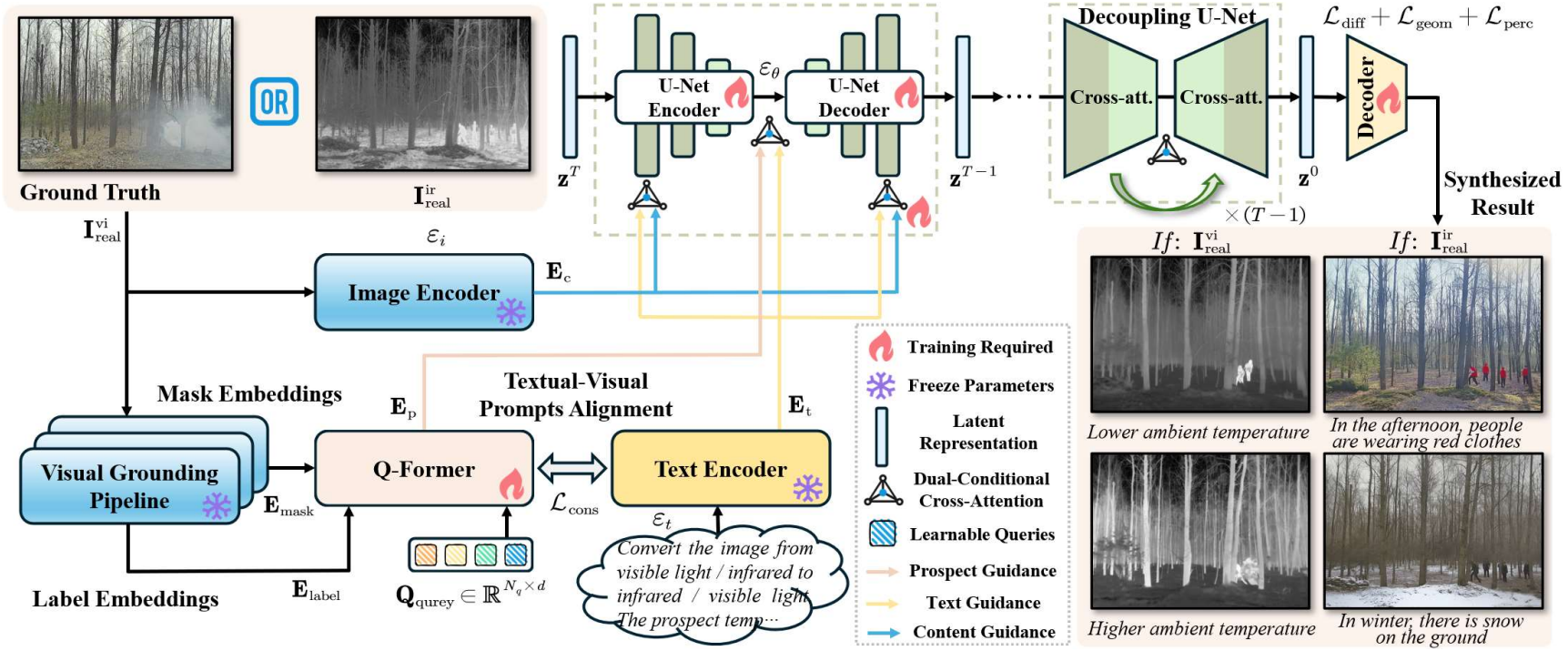



*For example, in the same visible scene, a car's temperature distribution may change significantly with its motion state, while a single infrared image may correspond to multiple visible appearances or environmental conditions.*

# [Method]

*We formulate image translation as a vision-language guided denoising diffusion process, enabling flexible conditioning and open-world generalization.*



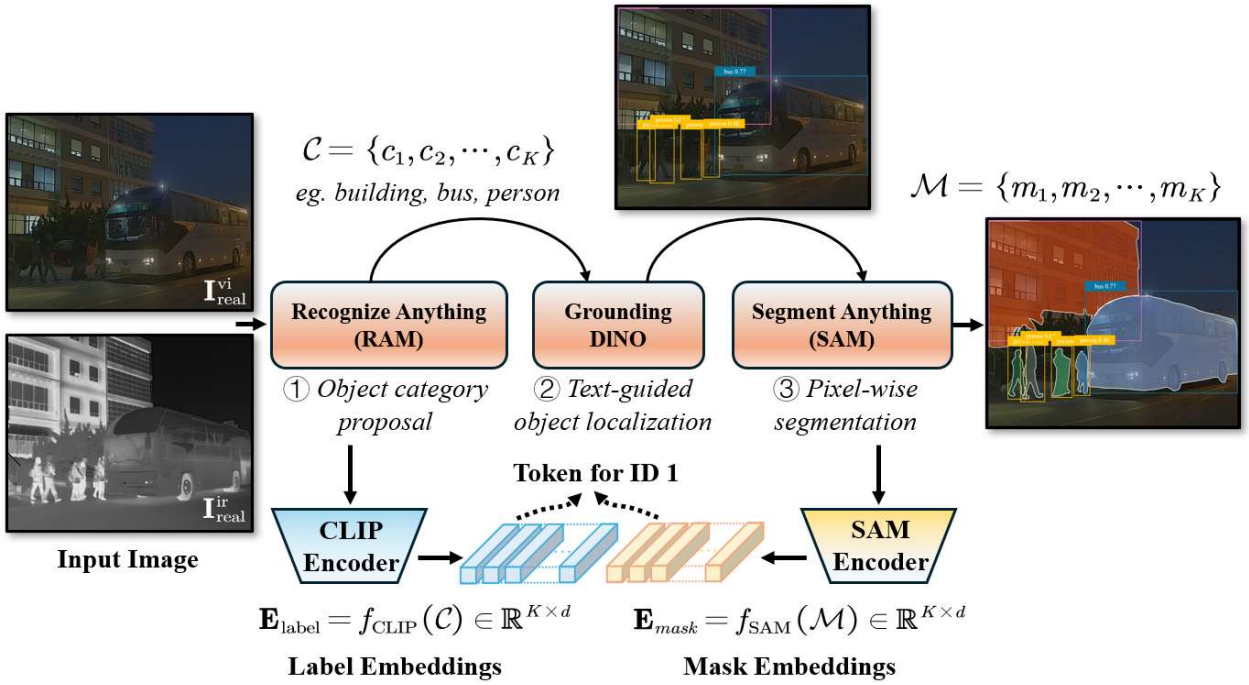

- We present SynthRGB-T, a novel multimodal image synthesis framework that leverages diffusion models and vision–language prompts to generate high-fidelity, diverse, and controllable images.
- We design a Visual Grounding Pipeline (VGP) that harnesses the world knowledge of foundation models to derive visually aligned and semantically coherent guidance. In addition, a Decoupled Injection Mechanism (DIM) is introduced to selectively inject multiple conditions during the fusion stage, mitigating semantic conflicts and information interference.
- We design a Dual-Conditional Cross-Attention (DCCA) module that enhances multimodal fusion by promoting collaborative representation learning across modality embeddings in the latent feature space.

# [Visual Grounding Pipeline]

- A pre-trained VGP is employed to assist in the early extraction of region-of-interest features during training
- This pipeline integrates three sequential foundation models—Recognize Anything (RAM), Grounding DINO, and Segment Anything (SAM)—to progressively identify, localize, and segment the foreground regions.
- A three-branch Q-Former architecture is designed to associate the region-of-interest mask, label information, and textual prompt within a unified representation space. The inputs to the Q-Former are paired foreground embeddings extracted by VGP, focusing on object level feature extraction to generate independent foreground tokens.

$$\mathbf{E}_{\mathrm{p}} = f_{\mathrm{Q-Former}}\left(\mathbf{E}_{\mathrm{mask}}, \mathbf{E}_{\mathrm{label}}, \mathbf{Q}_{\mathrm{query}}\right)$$

*Beyond automated dense object labeling, the text description and corresponding mask of each foreground are fed into the CLIP encoder and SAM encoder to establish a one-to-one alignment between textual and visual representations. This enables the VGP to generate conditional embeddings mask and label for each object in a zero-shot manner, thereby constructing an implicit translation prior*



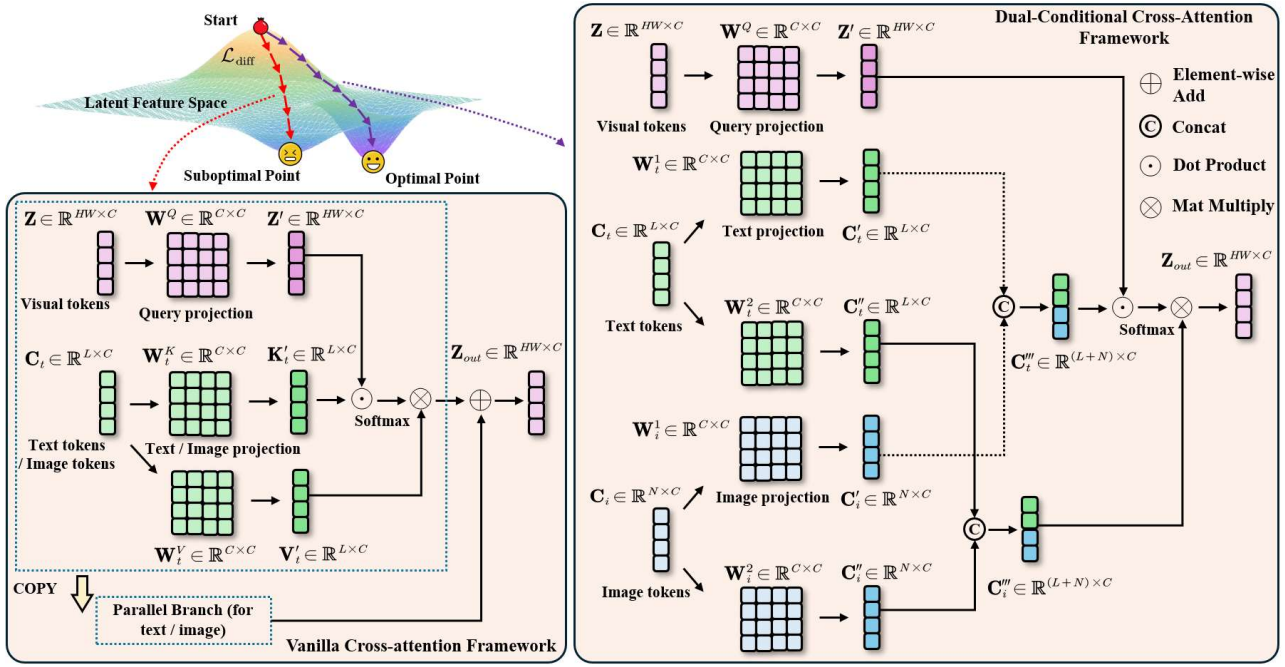
# [Dual-Conditional Cross-Attention & Decoupled Injection]

- To enable the denoising U-Net to support the decoupled injection mechanism, we designed a DCCA guided by joint text-content or text-foreground.
- In DCCA, we introduce two sets of trainable linear projection layers to handle image features and text features separately. Unlike previous methods that apply separate cross-attentions for each modality, DCCA concatenates the keys and values of multi-source features, then employs the U-Net query feature to initiate a unified cross-attention.

$$\mathbf{K} = \text{Concat}\left(\mathbf{C}_i\mathbf{W}_i^1, \mathbf{C}_t\mathbf{W}_t^1\right)$$

$$\mathbf{V} = \text{Concat}\left(\mathbf{C}_i\mathbf{W}_i^2, \mathbf{C}_t\mathbf{W}_t^2\right)$$

$$\mathbf{Z}_{\text{out}} = \text{Softmax}\left(\frac{\left(\mathbf{Z}\mathbf{W}^Q\right)\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}$$



# [Loss Function]

- To ensure alignment between the query token extracted by VGP and the predefined prompt embedding throughout the diffusion process, a similarity constraint is applied using MSE loss and cosine similarity loss.
- After aligning the visual and text prompts, the second stage of training is guided by three losses: diffusion loss to ensure text consistency, geometric loss to preserve content structure, and perceptual constraint loss to enhance visual-semantic fidelity.

$$\mathcal{L}_{\text{cons}} = \lambda_1 \underbrace{\|\mathbf{E}_{\text{t}} - \mathbf{E}_{\text{p}}\|^2}_{\mathcal{L}_{\text{mse}}} + \lambda_2 \underbrace{\left(1 - \frac{\mathbf{E}_{\text{t}} \cdot \mathbf{E}_{\text{p}}}{\|\mathbf{E}_{\text{t}}\| \, \|\mathbf{E}_{\text{p}}\|}\right)}_{\mathcal{L}_{\text{cos}}}$$

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z},\mathbf{E},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, t, [\mathbf{E}_{\text{c}}, \mathbf{E}_{\text{p}}, \mathbf{E}_{\text{t}}]\right)\|_2^2\right]$$

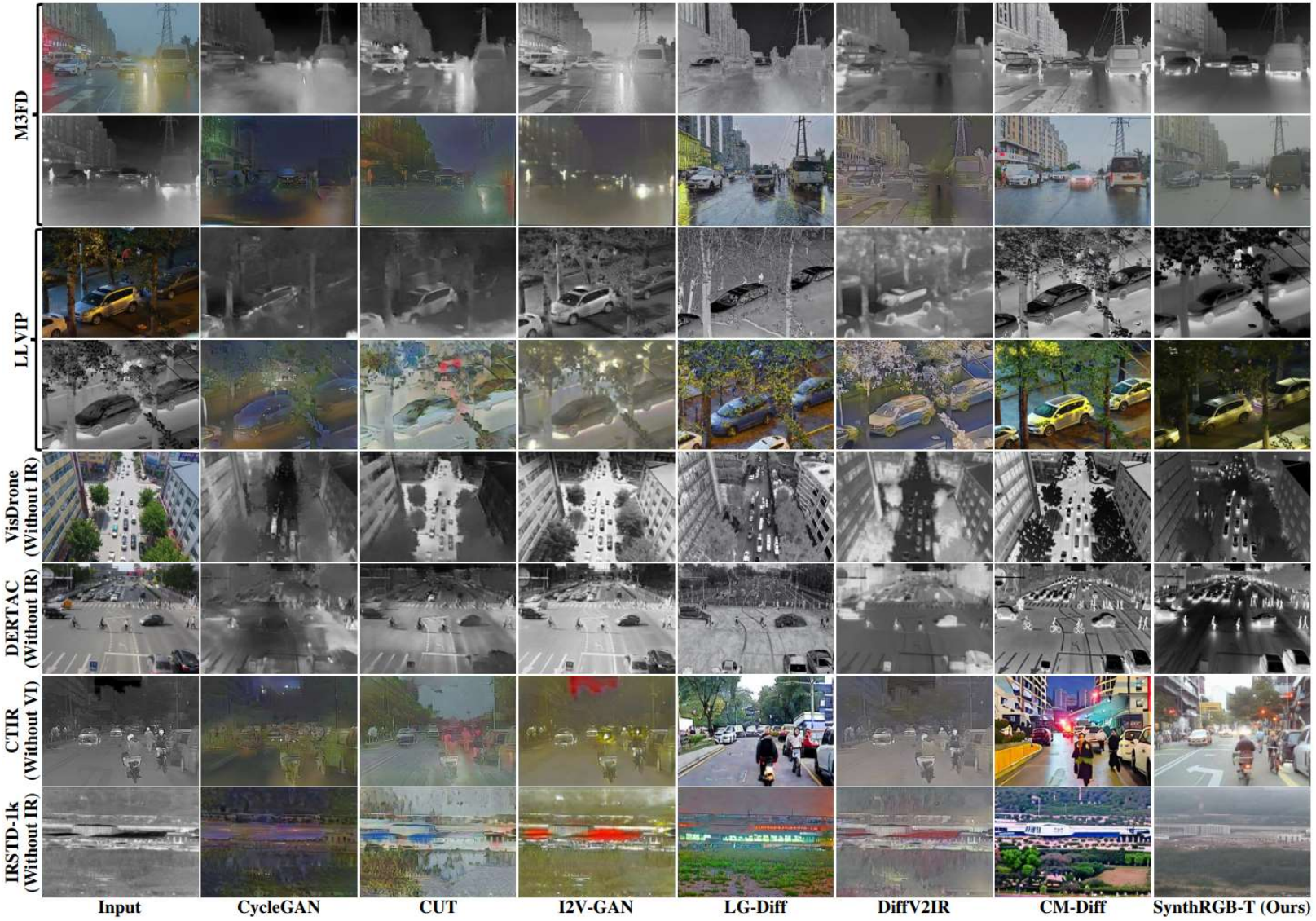
$$\mathcal{L}_{\text{geom}} = 1 - \frac{\left\langle g(\hat{\mathbf{I}}), g\left(\mathbf{I}_{\text{src}}\right)\right\rangle}{\|g(\hat{\mathbf{I}})\|_2 \cdot \|g\left(\mathbf{I}_{\text{src}}\right)\|_2 + \epsilon}$$

$$\mathcal{L}_{\text{perc}} = \sum_i \left\|\phi_i(\hat{\mathbf{I}}) - \phi_i\left(\mathbf{I}_{\text{target}}\right)\right\|_1$$

# [Comparison with SOTA Methods]

- Qualitative results of multimodal translation tasks showing several representative SOTA approaches. The first four rows correspond to bidirectional translation on M3FD and LLVIP. The fifth and sixth rows display visual-only benchmarks (VisDrone, DERTAC). The seventh and eighth rows show infrared-only benchmarks (CTIR, IRSTD-1k).

- SynthRGB-T achieves the highest visual realism, particularly on unseen open benchmarks, enhancing the credibility of future synthetic datasets.



# [Comparison with SOTA Methods]

- Evaluations are conducted on five benchmarks: M3FD, LLVIP, AVMS, CTIR, and VisDrone, which include both paired and unpaired settings.
- Traditional GAN approaches tend to produce lower SSIM and higher LPIPS, indicating weaker cross-modal consistency due to limited pixel-level correspondence. In contrast, diffusion based methods achieve more stable performance but primarily model the macro-level target-domain style, without ensuring the true semantic alignment of translated pixels. Our method, achieves state-of-the-art results in both directions, effectively bridging the modality gap.

Table 1. Comparison of I2V task. Red, orange, and yellow indicate the top three results. ↑ means higher is better, ↓ means lower is better.

| Method | $M^3FD$ (Paired infrared → visible) | | | | LLVIP (Paired infrared → visible) | | | | AVMS (Paired infrared → visible) | | | | CTIR (Unpaired infrared → visible) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NIQE↓ | LPIPS↓ | FID↓ | SSIM↑ | NIQE↓ | LPIPS↓ | FID↓ | SSIM↑ | NIQE↓ | LPIPS↓ | FID↓ | SSIM↑ | NIQE↓ | LPIPS↓ | FID↓ | SSIM↑ |
| CycleGAN [61] | 8.92 | 0.317 | 132.4 | 0.404 | 8.45 | 0.229 | 127.6 | 0.503 | 9.01 | 0.323 | 136.8 | 0.398 | 8.77 | 0.305 | 129.2 | 0.409 |
| StegoGAN [52] | 6.34 | 0.245 | 95.3 | 0.517 | 6.10 | 0.294 | 89.8 | 0.533 | 6.49 | 0.257 | 98.1 | 0.521 | 6.22 | 0.241 | 91.2 | 0.526 |
| UNIT [22] | 8.81 | 0.286 | 118.5 | 0.471 | 8.25 | 0.273 | 112.7 | 0.512 | 8.59 | 0.291 | 120.4 | 0.469 | 8.36 | 0.276 | 114.9 | 0.478 |
| CUT [35] | 7.10 | 0.412 | 104.2 | 0.502 | 6.98 | 0.398 | 101.5 | 0.549 | 7.28 | 0.419 | 107.1 | 0.497 | 7.09 | 0.405 | 103.8 | 0.506 |
| I2V-GAN [19] | 4.55 | 0.202 | 208.1 | 0.561 | 5.60 | 0.187 | 98.4 | 0.578 | 5.87 | 0.196 | 105.3 | 0.566 | 5.74 | 0.191 | 101.1 | 0.572 |
| LG-Diff [5] | 5.96 | 0.118 | 78.3 | 0.654 | 4.33 | 0.108 | 42.8 | 0.703 | 4.60 | 0.120 | 58.3 | 0.678 | 4.96 | 0.129 | 46.3 | 0.692 |
| DiffV2IR [37] | 5.84 | 0.153 | 116.7 | 0.686 | 4.66 | 0.123 | 50.4 | 0.671 | 4.97 | 0.136 | 59.2 | 0.649 | 4.70 | 0.113 | 53.4 | 0.659 |
| CM-Diff [10] | 4.50 | 0.114 | 43.9 | 0.665 | 4.15 | 0.101 | 39.8 | 0.718 | 4.54 | 0.117 | 55.6 | 0.685 | 5.34 | 0.107 | 42.7 | 0.699 |
| **Ours** | **4.30** | **0.107** | **40.3** | **0.753** | **4.25** | **0.057** | **34.2** | **0.885** | **4.48** | **0.112** | **49.6** | **0.729** | **4.45** | **0.093** | **36.3** | **0.743** |

Table 2. Comparison of V2I task. Red, orange, and yellow indicate the top three results. ↑ means higher is better, ↓ means lower is better.

| Method | $M^3FD$ (Paired visible → infrared) | | | | LLVIP (Paired visible → infrared) | | | | AVMS (Paired visible → infrared) | | | | VisDrone (Unpaired visible → infrared) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NIQE↓ | LPIPS↓ | FID↓ | SSIM↑ | NIQE↓ | LPIPS↓ | FID↓ | SSIM↑ | NIQE↓ | LPIPS↓ | FID↓ | SSIM↑ | NIQE↓ | LPIPS↓ | FID↓ | SSIM↑ |
| CycleGAN [61] | 5.98 | 0.258 | 93.4 | 0.428 | 8.09 | 0.214 | 118.3 | 0.530 | 8.55 | 0.302 | 128.3 | 0.421 | 8.30 | 0.284 | 121.4 | 0.431 |
| StegoGAN [52] | 6.40 | 0.249 | 110.6 | 0.544 | 5.75 | 0.274 | 84.7 | 0.554 | 6.04 | 0.239 | 91.8 | 0.543 | 5.85 | 0.225 | 86.8 | 0.552 |
| UNIT [22] | 8.22 | 0.270 | 110.2 | 0.490 | 7.70 | 0.254 | 105.0 | 0.534 | 8.05 | 0.271 | 112.8 | 0.488 | 7.77 | 0.259 | 107.9 | 0.501 |
| CUT [35] | 6.60 | 0.383 | 97.9 | 0.523 | 6.50 | 0.374 | 94.4 | 0.568 | 6.78 | 0.392 | 99.6 | 0.518 | 6.64 | 0.374 | 96.5 | 0.528 |
| I2V-GAN [19] | 4.21 | 0.190 | 193.5 | 0.589 | 5.22 | 0.176 | 91.5 | 0.602 | 5.46 | 0.183 | 98.6 | 0.593 | 4.37 | 0.177 | 94.0 | 0.608 |
| LG-Diff [5] | 5.59 | 0.103 | 72.8 | 0.735 | 4.02 | 0.101 | 37.0 | 0.742 | 4.28 | 0.144 | 54.2 | 0.686 | 4.61 | 0.121 | 43.1 | 0.732 |
| DiffV2IR [37] | 4.15 | 0.092 | 39.5 | 0.689 | 4.74 | 0.089 | 35.9 | 0.774 | 4.23 | 0.112 | 51.7 | 0.736 | 4.20 | 0.099 | 37.8 | 0.748 |
| CM-Diff [10] | 4.28 | 0.125 | 40.8 | 0.626 | 4.95 | 0.094 | 40.0 | 0.751 | 4.10 | 0.100 | 48.7 | 0.715 | 4.97 | 0.091 | 39.7 | 0.725 |
| **Ours** | **4.00** | **0.078** | **37.5** | **0.783** | **3.95** | **0.053** | **31.8** | **0.922** | **4.17** | **0.095** | **46.1** | **0.754** | **4.14** | **0.086** | **33.8** | **0.772** |

# [Ablation Study]

- Each component contributes to the final performance. Removing VGP, DIM, DCCA, or text-visual alignment leads to performance degradation.
- DCCA and DIM improve multimodal guidance fusion. They reduce conflicts among text, foreground, and content conditions.
- The full SynthRGB-T achieves the best balance between fidelity and diversity. It follows prompts more accurately while preserving realistic scene structures.

Table 3. Ablation components. Red, orange, and yellow indicate the top three results. ↑ means higher is better, ↓ means lower is better.

| ID | VGP | DIM | DCCA | $\mathcal{L}_{\text{cons}}$ | NIQE↓ | LPIPS↓ | FID↓ | SSIM↑ |
|---|---|---|---|---|---|---|---|---|
| I | × | × | × | × | 7.96 | 0.356 | 144.2 | 0.407 |
| II | × | ✓ | ✓ | × | 7.41 | 0.323 | 128.0 | 0.439 |
| III | ✓ | ✓ | × | ✓ | 5.58 | 0.154 | 60.3 | 0.699 |
| IV | ✓ | × | ✓ | ✓ | 5.85 | 0.172 | 62.1 | 0.675 |
| V | ✓ | ✓ | ✓ | × | 6.62 | 0.248 | 88.6 | 0.629 |
| VI | ✓ | ✓ | ✓ | ✓ | 4.22 | 0.085 | 38.7 | 0.793 |

Table 4. Ablation on Diversity Generation. Red, orange, and yellow indicate the top three results. ↑ means higher is better, ↓ means lower is better.

| Method | infrared → visible | | visible → infrared | |
|---|---|---|---|---|
| | LPIPS↑ | FID↓ | LPIPS↑ | FID↓ |
| w/o DIM | 0.132±0.008 | 38.9±0.1 | 0.102±0.021 | 35.5±0.5 |
| w/o DCCA | 0.158±0.012 | 38.0±0.2 | 0.098±0.014 | 35.1±0.2 |
| w/o $\mathcal{L}_{\text{cons}}$ | 0.108±0.025 | 41.5±0.4 | 0.080±0.011 | 37.4±0.1 |
| SynthRGB-T | 0.179±0.013 | 36.9±0.2 | 0.126±0.008 | 34.5±0.1 |

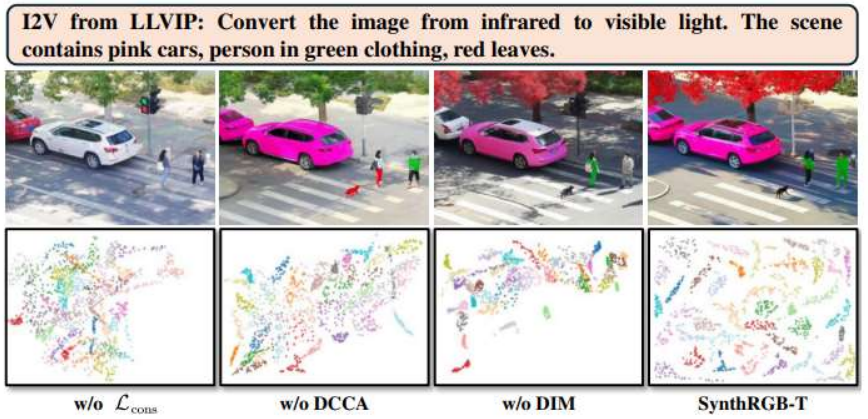
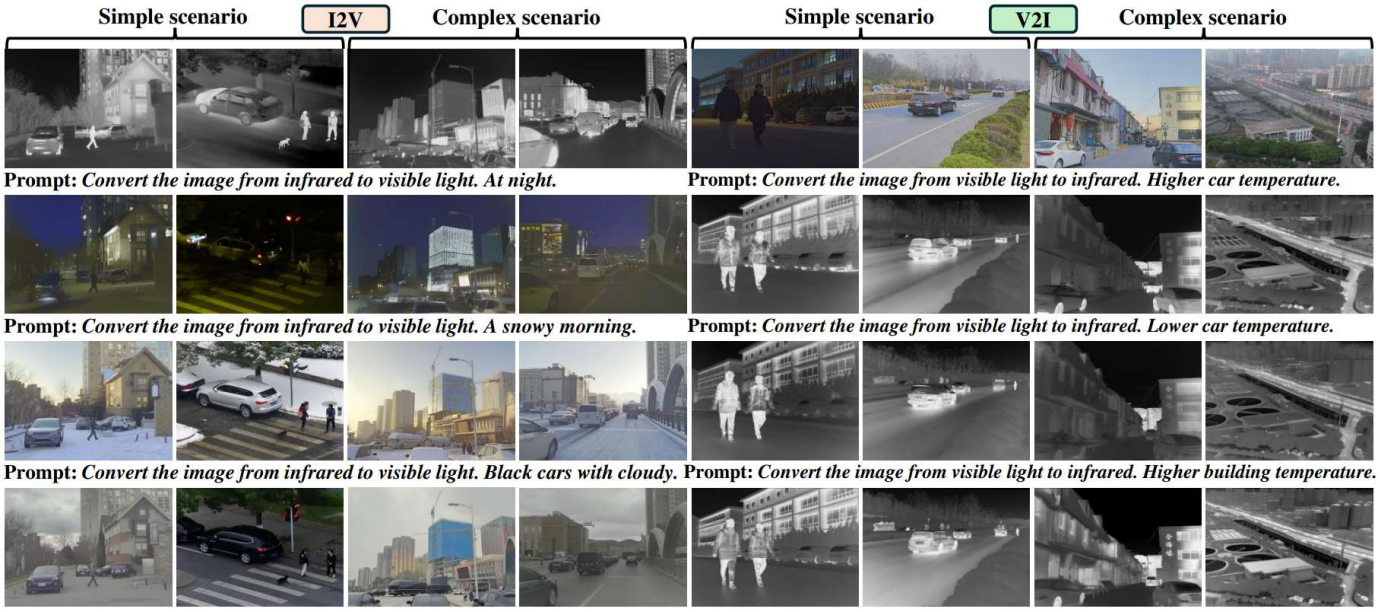
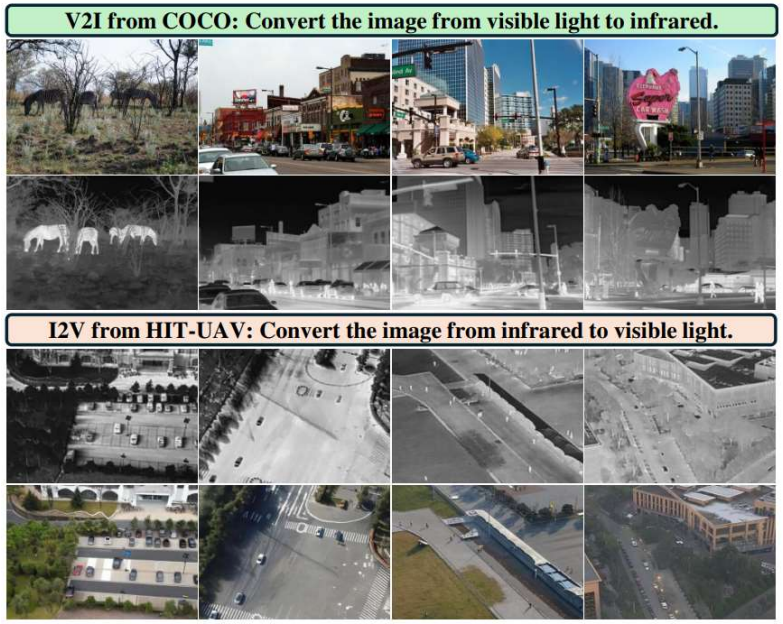



Figure 8. Qualitative results of ablation study.

# [Unpaired Translation Test]

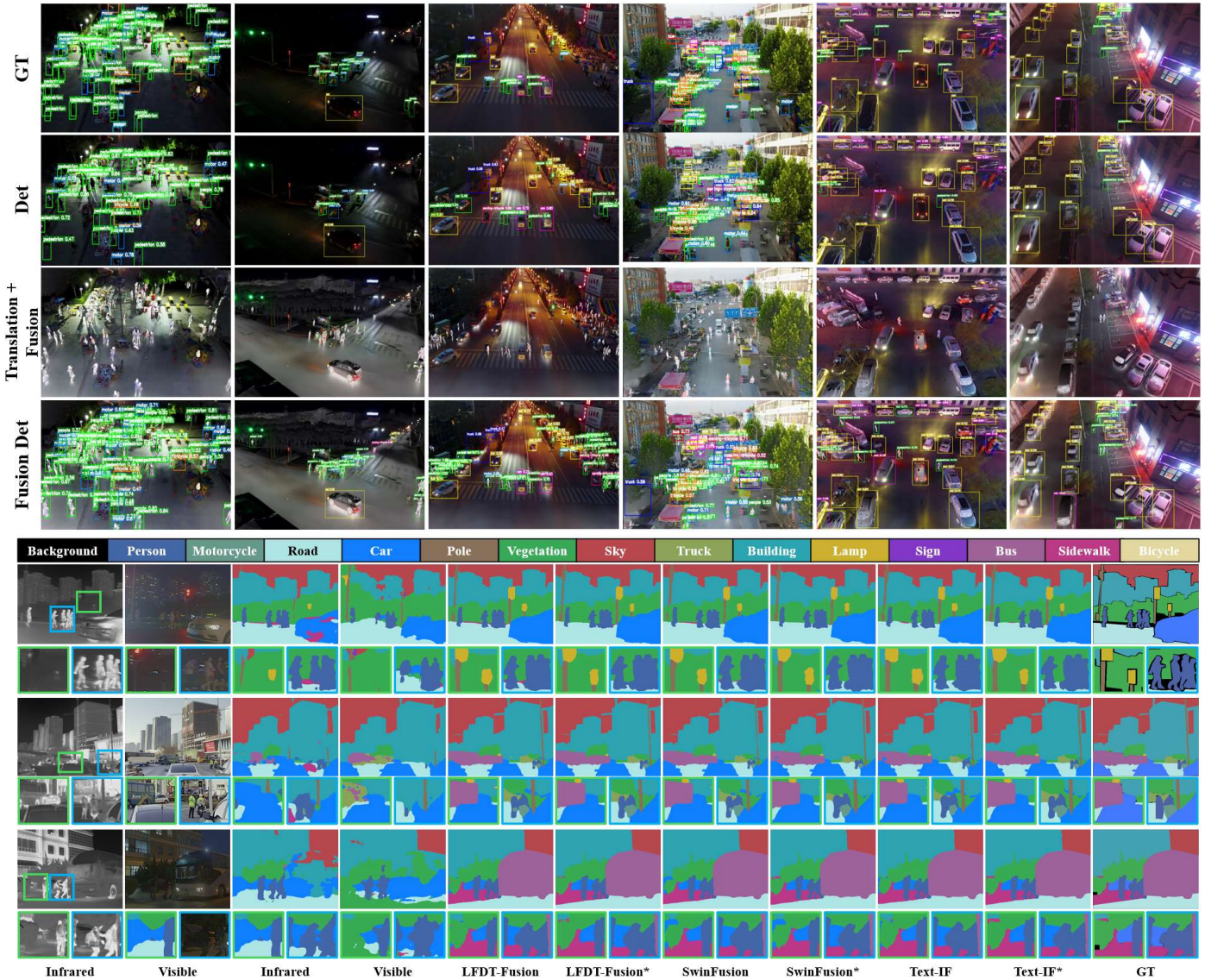
- For both the I2V and V2I tasks, eight real-world scenes were selected from public benchmarks. SynthRGB-T is capable of synthesizing visually distinct yet semantically consistent visible and thermal counterparts, driven by user-defined prompts. The generated results exhibit high photorealism and finegrained structural fidelity. Furthermore, the prompts can be expressed in diverse linguistic forms and styles, enabling flexible and intuitive control over the translation process.



# [Adaptability Verification for High-Level Tasks]

- Compared with direct training, in-corporating additional synthetic data leads to noticeable improvements in image detail preservation, edge sharpness, and object recognition accuracy. For object detection, the fused images exhibit more precise boundary localization of salient objects and achieve detections with higher confidence. For semantic segmentation, the generated maps clearly distinguish different class regions and show close alignment with the ground truth. These observations demonstrate that synthetic data can effectively enhance downstream tasks by providing finer details and achieving a better balance of contextual information.



CVPR
JUNE 3-7, 2026
DENVER
COLORADO

# Thanks for listening!

CHANG' AN UNIVERSITY
長安大學
CHANG' AN UNIVERSITY

THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學