OMNIRSI: TAMING LATENT DIFFUSION FOR UNIFIED MULTIMODAL REMOTE SENSING IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal remote sensing image generation emerges as a promising paradigm for recovering complete and complementary observations from heterogeneous and partially observed remote sensing data. However, existing methods predominantly depend on sensor-specific fine-tuning of natural image generation models, limiting unified multimodal generation capacity due to the absence of shared representations and robust conditioning mechanisms. To address these challenges, we propose OmniRSI, a unified multimodal remote sensing image generation model. Specifically, we develop a remote sensing-specific variational autoencoder that captures multimodal priors and promotes a locally smooth latent space for unified representation and robust reconstruction across diverse observed data. Building upon that, we further fully finetune Stable Diffusion 3.5 with a ControlNet branch to support unified multimodal generation and translation conditioned on textual and visual cues. Furthermore, we devise a training-free cyclic latent alignment mechanism that iteratively enriches incomplete visual conditions with generated multimodal information, progressively improving guidance reliability while preserving observed measurements. In addition, we construct RSI-MM, a large-scale dataset comprising 11.5 million multi-resolution global images with text annotations and partially paired observations, to enable training and evaluation. Extensive experiments verify that OmniRSI achieves superior performance in multimodal reconstruction, generation and translation across diverse scenarios.

1 INTRODUCTION

Multimodal remote sensing images are fundamental carriers for earth observation, where different modalities capture distinct yet complementary information of earth surface. Contemporary satellites are typically equipped with modality-specific sensors that acquire diverse observations, including panchromatic (PAN), multispectral (MSI), and synthetic aperture radar (SAR) imagery. These heterogeneous observations support a wide range of applications Wu et al. (2025). However, due to sensor-specific acquisition mechanisms, orbital constraints, and varying observation conditions, complete and spatially aligned multimodal observations are often unavailable in practice. This scarcity of paired multimodal data has motivated the development of generative models Shao et al. (2025); Yu et al. (2026) for remote sensing image synthesis and completion.

Generative models Hu et al. (2026) aim to learn underlying data distributions and synthesize novel samples with similar characteristics. Recent advances in diffusion models have pushed the frontiers of visual generation tasks due to their remarkable performance He et al. (2025); Xiao et al. (2025); Esser et al. (2024). Most multimodal remote sensing image generation approaches follow the latent diffusion-based generative pipeline Zhao et al. (2026); Tang et al. (2025); Yu et al. (2024). It generally consists of three essential stages: representation, generation, and translation. At the first stage, some methods Zhang et al. (2025) directly utilize frozen variational autoencoders (VAEs) pretrained on natural images Rombach et al. (2022), which provide efficient latent compression but yield suboptimal representations for remote sensing data. Other methods Chen et al. (2026) explore more compatible solutions by performing modality-specific VAE adaptation. At the second stage, most methods perform generative sampling in the latent space by gradually perturbing data through the forward diffusion process and learning a reverse denoising process to recover the underlying data distribution Fuest et al. (2026). They typically adapt large-scale pretrained diffusion models via parameter-efficient fine-tuning to efficiently transfer general visual priors to remote sensing do-

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

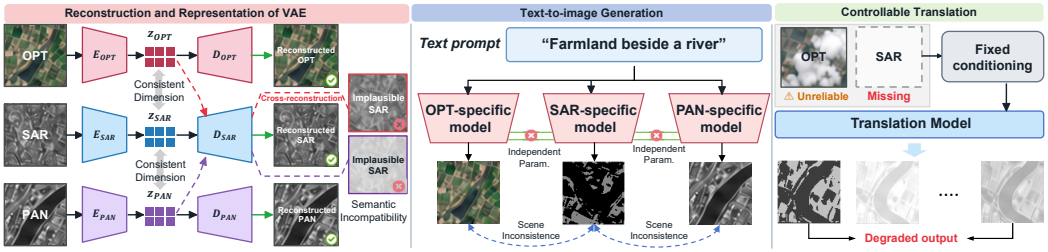

Figure 1: Overview and limitations of latent diffusion-based remote sensing image generation.

mains while retaining their inherent text-to-image (T2I) generation capabilities. However, these methods are usually developed for individual modalities, such as text-to-RGB Liu et al. (2025); Zhang et al. (2024a) and text-to-SAR generation Liu et al. (2026); Fan et al. (2025). At the third stage, auxiliary modality conditions are incorporated into generative models to guide the sampling process. These conditions are encoded and injected into the generation backbone through dedicated branches or attention-based mechanisms Ye et al. (2026); Zhang et al. (2026b). Such extensions rely on predefined paired relationships, and can thus be viewed as task-specific visual translation, where a conditional distribution is learned from a source modality to a target modality.

Despite these advances, existing methods remain insufficient for multimodal remote sensing generation, exhibiting three fundamental limitations: semantically misaligned latent spaces, fragmented multimodal generation, and rigid conditional guidance (Fig. 1). First, modality-specific VAE adaptation primarily ensures dimensional consistency while leaving substantial semantic discrepancies across modalities unresolved and limiting the representation transferability. Second, current generative models are typically adapted to individual modalities, preventing effective sharing of generative priors across heterogeneous observations and resulting in fragmented generation capabilities. Third, conditional translation methods rely on predefined source-target relationships and fixed guidance throughout inference, limiting the flexibility across modality combinations. When source observations are incomplete or uncertain, such static conditioning cannot adaptively refine unreliable guidance, making the generation process vulnerable to accumulated errors. More fundamentally, these components are largely developed in isolation, lacking an integrated design that jointly couples multimodal representation learning, unified generation, and adaptive condition refinement.

To address these challenges, we propose OmniRSI, a unified multimodal remote sensing generation model that tightly couples the representation, generation and translation. Specifically, we organize heterogeneous observations into modality-aware slots according to sensor-specific characteristics and develop a remote sensing-specific variational autoencoder (RS-VAE) to learn shared latent representations from incomplete and heterogeneous observations. By combining masked reconstruction with latent smoothness regularizations, it is able to capture cross-modal dependencies and promotes a locally smooth latent space, thereby enabling robust representation under incomplete observations. Building upon this, we adapt Stable Diffusion 3.5 with a ControlNet branch to achieve unified multimodal generation from textual descriptions and conditional translation from partially available observations. To improve the condition reliability and alleviate accumulated errors, we introduce a training-free cyclic latent alignment (CLA) mechanism that iteratively enriches incomplete visual conditions with generated multimodal information while preserving observed measurements during inference. To support training and evaluation, we construct RSI-MM, a large-scale multimodal remote sensing dataset containing 11.5 million multi-resolution global images with text annotations and partially paired observations. Extensive experiments demonstrate that OmniRSI consistently outperforms the state-of-the-art methods across reconstruction, generation, and translation tasks.

Our main contributions can be summarized as follows: 1) We develop the RS-VAE that learns a unified latent representation from incomplete and heterogeneous observations by leveraging modality-aware slots and tailored reconstruction strategies. 2) We fully fine-tune Stable Diffusion 3.5 with a ControlNet branch to tightly couple multimodal generation and translation. Furthermore, a training-free sampling mechanism is devised to progressively refine unreliable conditions while preserving observed measurements. 3) We construct RSI-MM, a large-scale multimodal remote sensing benchmark with diverse global observations, text descriptions, and partially paired modalities, facilitating comprehensive training and evaluation for unified multimodal learning.

2 Related Work

Generative Models have achieved remarkable progress in visual synthesis, mainly represented by autoregressive and diffusion-based approaches. Autoregressive models learn image distributions by predicting discrete visual tokens produced by pretrained tokenizers. Recent advances have explored both improved token representations and flexible prediction paradigms. For the former, RQ-VAE Lee et al. (2022) introduces recursive residual quantization to reduce mapping errors, FSQ Mentzer et al. (2024) directly quantizes continuous features, and TokenFlow Qu et al. (2025) explores multi-codebook representation to capture visual semantics at different granularities. For the latter, recent methods Liu et al. (2024b); Wang et al. (2025b); Pang et al. (2026b) relax the fixed raster-order prediction paradigm through flexible token grouping, parallel prediction, and patch-level token aggregation, while VAR Tian et al. (2024) further introduces next-scale prediction for coarse-to-fine image synthesis. Diffusion models generate samples by progressively perturbing data and learning the reverse inference process. Latent Diffusion Rombach et al. (2022) improves scalability by performing diffusion in a compressed latent space, enabling large-scale generative systems such as Stable Diffusion (SD) series Esser et al. (2024); Podell et al. (2024). Beyond text conditioning, recent studies have introduced more flexible control mechanisms. Training-free methods, including classifier guidance Dhariwal & Nichol (2021), classifier-free guidance Gao et al. (2026); Lin et al. (2026), and inversion-based editing methods Xie et al. (2026), manipulate sampling trajectories without additional training. Meanwhile, fine-tuning methods Hu et al. (2025) such as T2I-Adapter Mou et al. (2024) and ControlNet Zhang et al. (2023); Motorcu & Cetin (2026) incorporate external conditions by lightweight adapters or additional control branches.

Remote sensing image generation predominantly follows diffusion-based generation paradigms and can be categorized into text-conditioned and visual-conditioned generation. For text-conditioned generation, several studies investigate modality-specific remote sensing image synthesis. Txt2Img-MHN Xu et al. (2023b) utilizes hierarchical prototype learning to improve semantic alignment and scene consistency for RGB generation, while Text2Earth Liu et al. (2025) constructs a large-scale dataset and develops a model for generating diverse optical scenes with varying spatial resolutions. HSIGene Pang et al. (2026a) introduces spectral-aware generation strategies and multi-condition controllability for hyperspectral image generation. HuiYanEarth-SAR Liu et al. (2026) explores text-guided SAR image generation by incorporating SAR-specific imaging characteristics and specific priors. Beyond textual conditions, recent studies introduce additional visual or structural guidance to improve controllability and spatial consistency. Any2RSI Zhang et al. (2026b) explores controllable generation by incorporating diverse control signals and enriched descriptions. DSP Bao et al. (2026) investigates layout-to-image generation under limited data regimes by disentangling semantic concepts and visual primitives. CC-Diff Zhang et al. (2024a), AeroGen Tang et al. (2025), and OFDiff Ye et al. (2026) incorporate geographic or observation-specific conditions to improve the fidelity and enable region-aware controllability. However, these approaches generally focus on modality-specific generation with predefined conditioning, leaving unified generation across heterogeneous modalities from partially observed multimodal inputs still unexplored.

3 Method

OmniRSI aims to establish a unified model for multimodal remote sensing generation under heterogeneous and partially observed data (Fig. 2). It is conceptualized into three vital functions, including multimodal image reconstruction, text-to-image generation and image-to-image translation.

Dataset Construction. To fully support these tasks, we construct RSI-MM that contains 11.5 million multi-resolution images collected from diverse public remote sensing datasets, including EuroSAT Helber et al. (2018), GID Tong et al. (2020), Git-10M Liu et al. (2025), OSCD Daudt et al. (2018), M3M-CR Xu et al. (2023a), OpenEarthMap-SAR Xia et al. (2025), PIE-RGB-SAR Zhang et al. (2024b), Inria Maggiori et al. (2017), WHU-OPT-SAR Li et al. (2022), SARLO Debuysère et al. (2026), RSCC Chen et al. (2025), RSICD Lu et al. (2017), M4-SAR Wang et al. (2025a), PSBench Wang et al. (2026b), FSARCap Zhang et al. (2026a), ChatEarthNet Yuan et al. (2024). As reported in Tab. 1, these datasets cover heterogeneous sensors, spatial resolutions, and modality configurations, including RGB, MSI, SAR, and PAN. For datasets without text annotations, we select the representative bands to form visualizations and feed them to GPT-5.4-mini Singh et al. (2025) with a specific prompt to generate textual descriptions of scene information (See Appendix A).

Table 1: Details of RSI-MM datasets. ✓: available; –: unavailable; ❍: supplemented by us.

Datasets	Source&Platform	Modality RGB	MSI	SAR	PAN	Text	Text Num.	Spat. Res.(m)	Band. Configs.	Image Num.	Image Size
Git-10M	GoogleEarth&Public datasets	✓	–	–	–	✓	10,452,649	0.5∼128	3	10,452,649	256 × 256
EuroSAT	Sentinel-2	✓	✓	–	–	❍	27,000	10/20/60	13	27,000	64 × 64
OSCD	Sentinel-2	✓	✓	–	–	❍	48	10/20/60	13	48	241∼1180
M3M-CR	PlanetScope&Sentinel-1	✓	✓	✓	–	❍	63,000	3/10	5	63,000	300/90
OpenEarthMap-SAR	UmbraSpace&NAIP&IGN&GSI	✓	–	✓	–	❍	5,033	0.15-0.5	4	5,033	1024 × 1024
PIE-RGB-SAR	Google Satellite&GF-3	✓	–	✓	–	❍	4,865	0.5-3	4	4,865	256 × 256
WHU-OPT-SAR	GF-1&GF-3	✓	✓	✓	–	✓	4,900	2/5	5	4,900	512× 512
SARLO	Umbra&Google Satellite	✓	–	✓	–	✓	358,698	0.8	4	119,566	1024 × 1024
RSCC	MAXAR	✓	–	–	–	✓	62,351	0.3∼0.8	3	62,351	512 × 512
RSICD	GoogleEarth&Baidu&MapABC	✓	–	–	–	✓	54,605	–	3	10,921	224 × 224
PSBench	GF&GE&WV&CBERS&Landsat	✓	✓	–	✓	❍	449,936	0.65∼30	5/8/9/11	449,936	256/512/1024
FSARCap	GF3	–	–	✓	–	✓	72,400	1.0∼5.0	1	14,480	1024 × 1024
GID	GF-2	✓	✓	–	–	✓	25,350	4.0	4	25,350	512× 512
Inria	–	✓	–	–	–	✓	1,800	0.3	3	1,800	512 × 512
M4-SAR	Sentinel-1&Sentinel-2	✓	–	✓	–	❍	112,174	10/60	4	112,174	512 × 512
ChatEarthNet	Sentinel-2	✓	✓	–	–	✓	169,220	10	9	164,656	256 × 256
RSI-MM(Total)	Above datasets	✓	✓	✓	✓	✓	11,864,029	0.15∼128	18	11,518,729	64∼1180

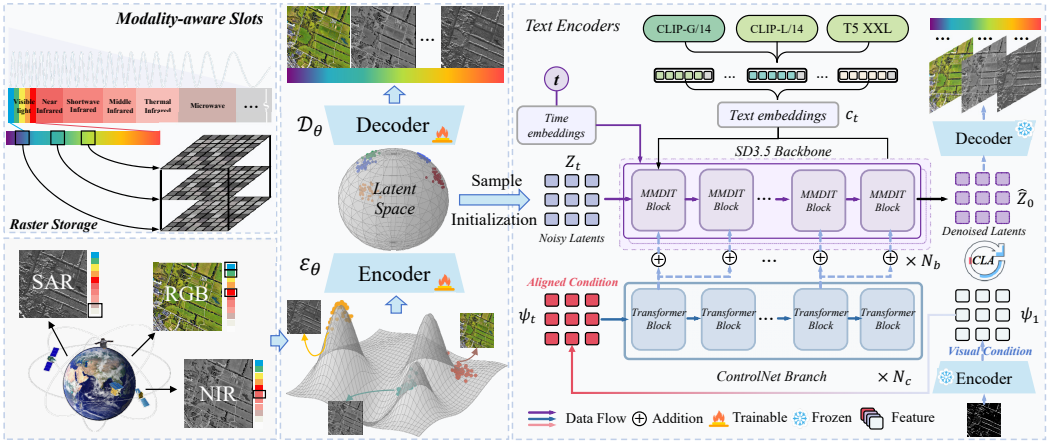

Figure 2: Overview of OmniRSI. Multimodal observations are slot-organized and encoded into a unified latent space. SD3.5 backbone is fine-tuned for generation, with a ControlNet branch injecting hierarchical condition guidance. The derived latents are decoded for multimodal reconstruction.

Multimodal Latent Representation Learning. To accommodate heterogeneous and partially observed remote sensing data, we organize observations into K predefined modality-aware slots according to their spectral characteristics and modality types (See Appendix B). Let $\mathbf{X} = \{\mathbf{x}_k\}_{k=1}^{K}$ denote the slot-organized data and $\mathbf{M} = \{\mathbf{m}_k\}_{k=1}^{K}$ their binary mask, indicating the slot availability. The valid input is $\mathbf{X}_v = \mathbf{M} \odot \mathbf{X}$, with $\mathbf{M}$ broadcast over spatial dimensions and $\odot$ denoting element-wise multiplication. This organization provides a consistent data dimension across different sensors and modality combinations. Building on this scheme, we develop RS-VAE with an encoder $\mathcal{E}_\theta$ that maps $\mathbf{X}$ to a latent variable $\mathbf{Z}$ and defines the approximate posterior $q_\theta(\mathbf{Z}\,|\,\mathbf{X})$ and a decoder $\mathcal{D}_\theta$ reconstructing the slot-organized observations. It retains the standard VAE architecture Esser et al. (2024) while adapting its input and output channels to the slot capacity K. To learn complementary multimodal representations and support reconstruction from incomplete observations, we jointly employ observed-slot and masked reconstruction constraints, with respect to $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{mask}}$:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{q_\theta(\mathbf{Z}|\mathbf{X}_v)}\Big[\|\mathbf{M}\odot(\mathcal{D}_\theta(\mathbf{Z})-\mathbf{X}_v)\|_2^2\Big], \quad \mathcal{L}_{\text{mask}} = \mathbb{E}_{q_\theta(\mathbf{Z}|\mathbf{X}_r)}\Big[\|\mathbf{M}\odot(\mathcal{D}_\theta(\mathbf{Z})-\mathbf{X}_v)\|_2^2\Big], \tag{1}$$

where $\mathbf{X}_r$ is obtained by randomly masking entire slots or spatial regions within individual slots. $\mathcal{L}_{\text{rec}}$ captures and internalizes cross-modal dependencies by jointly reconstructing available observations, while $\mathcal{L}_{\text{mask}}$ encourages recovery of withheld measurements from complementary modalities and spatial-spectral context. Furthermore, we promote locally smooth decoding by penalizing the sensitivity of $\mathcal{D}_\theta$ to symmetric latent perturbations:

$$\mathcal{L}_{\text{smooth}} = \mathbb{E}_{\mathbf{Z},\boldsymbol{\eta}}\Big[\|D_\theta(\mathbf{Z}+\boldsymbol{\eta}) - D_\theta(\mathbf{Z}-\boldsymbol{\eta})\|_1/(\|\boldsymbol{\eta}\|_1+\delta)\Big], \tag{2}$$

where $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma_\eta^2 \boldsymbol{I})$ is a small Gaussian perturbation and δ is a small positive value that ensures numerical stability. This constraint discourages abrupt changes in reconstruction under local latent

variations. Together with the posterior regularization $\mathcal{L}_{\mathrm{KL}} = \mathbb{E}_{q_\theta(\mathbf{Z}|\mathbf{X})}[D_{\mathrm{KL}}(q_\theta(\mathbf{Z} \mid \mathbf{X}) \| \mathcal{N}(0, \boldsymbol{I}))]$ that constrains representation toward an isotropic Gaussian prior, the overall objective is defined as:

$$\mathcal{L}_{\mathrm{vae}} = \mathcal{L}_{\mathrm{rec}} + \alpha \mathcal{L}_{\mathrm{mask}} + \gamma \mathcal{L}_{\mathrm{smooth}} + \beta \mathcal{L}_{\mathrm{KL}}, \tag{3}$$

where α, γ, and β balance each constraint. These four complementary objectives establish a shared representation space for unified multimodal generation and translation.

Multimodal Generation and Translation. Grounded in the unified latent space, we adopt SD3.5 Medium with N_b MMDiT blocks as the backbone for text-to-image generation (See Appendix C). Specifically, text prompts are jointly encoded by CLIP-G/14, CLIP-L/14 Beaumont (2022), and T5-XXL Raffel et al. (2020), collectively yielding the textual condition c. Meanwhile, the pretrained encoder $\mathcal{E}_\theta$ is frozen and maps the available data to a target representation $\mathbf{Z}_0 = \mathcal{E}_\theta(\mathbf{X})$. Following the Flow matching Lipman et al. (2023); Liu (2022) paradigm, given $\mathbf{Z}_0$ and independent Gaussian noise $\mathbf{Z}_1 \sim \mathcal{N}(0, \boldsymbol{I})$, an intermediate latent state $\mathbf{Z}_t$ is constructed as $\mathbf{Z}_t = (1 - \sigma_t)\mathbf{Z}_0 + \sigma_t \mathbf{Z}_1$, where $\sigma_t \in [0, 1]$ specifies the interpolation schedule. The model learns a conditional velocity field $v_\theta(\mathbf{Z}_t, t, c)$ by regressing the sample-wise target velocity:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\mathbf{Z}_0, \mathbf{Z}_1, t, c}\left[\| v_\theta(\mathbf{Z}_t, t, c) - \dot{\sigma}_t(\mathbf{Z}_1 - \mathbf{Z}_0) \|_2^2 \right], \tag{4}$$

where $\dot{\sigma}_t = d\sigma_t / dt$. At inference, one can utilize the classifier-free guidance to steer generative trajectories by combining unconditional and conditional velocity predictions:

$$v_{\mathrm{cfg}}(\mathbf{Z}_t, t, c) = v_\theta(\mathbf{Z}_t, t, \varnothing) + w\left[v_\theta(\mathbf{Z}_t, t, c) - v_\theta(\mathbf{Z}_t, t, \varnothing) \right], \tag{5}$$

where $\varnothing$ denotes the null text condition and w controls the guidance strength. Starting from stochastic noise $\mathbf{Z}_1 \sim \mathcal{N}(0, \boldsymbol{I})$, an ordinary differential equation solver integrates $d\mathbf{Z}_t/dt = v_{\mathrm{cfg}}(\mathbf{Z}_t, t, c)$ from $t = 1$ to $t = 0$ to obtain the generated latent $\hat{\mathbf{Z}}_0$. Then the frozen decoder $\mathcal{D}_\theta$ maps back to the pixel space as $\hat{\mathbf{X}} = \mathcal{D}_\theta(\hat{\mathbf{Z}}_0)$, producing full-slot outputs across heterogeneous modalities.

To support conditional translation from observed data, we augment the backbone with a trainable ControlNet branch containing N_c Transformer blocks. The observed data $\mathbf{X}$ are first projected into the latent space to obtain the condition as $\boldsymbol{\psi} = \mathcal{E}_\theta(\mathbf{X})$, which is then processed by the ControlNet branch to produce hierarchical control features $\{\mathbf{h}_j^c\}_{j=1}^{N_c}$. These features are residually added to the hidden states $\mathbf{H}_i$ of the SD3.5 backbone after each MMDiT block, which can be expressed as:

$$\mathbf{H}_i \leftarrow \mathbf{H}_i + \mathbf{h}_{g(i)}^c, \qquad g(i) = 1 + \lfloor N_c \cdot (i-1)/N_b \rfloor, \quad i = 1, \ldots, N_b, \tag{6}$$

where $\lfloor \cdot \rfloor$ denotes the floor operation. For $N_c \leq N_b$, this assignment distributes control features across consecutive groups of backbone blocks, providing visual guidance at multiple depths. Accordingly, each control feature modulates a group of consecutive MMDiT blocks, allowing the partially observed modalities to provide hierarchical visual guidance throughout generation. Moreover, both of RS-VAE and SD3.5 backbone are frozen, while the additional branch is optimized with the similar objective in Eq. (4) using the conditional velocity $v_\theta(\mathbf{Z}_t, t, c, \boldsymbol{\psi})$. This formulation supports translation from different subsets of observed modalities within a shared generative model.

Cyclic Latent Alignment Textual descriptions primarily convey global semantics but remain inherently sparse for wide-field-of-view remote sensing images, where fine-grained spatial and spectral details resist exhaustive specification. Partially observed modalities offer more explicit visual guidance, but their incompleteness may bias conditional velocity predictions. Given partially observed data $\mathbf{X}$ with availability mask $\mathbf{M}$, the coarse visual condition is obtained by $\boldsymbol{\psi}_1 = \mathcal{E}_\theta(\mathbf{M} \odot \mathbf{X})$. The Euler solver is then used to approximate the solution, with the forward and reverse steps given by:

$$\textbf{Fwd: } \mathbf{Z}_{t+\Delta t} = \mathbf{Z}_t + v(\mathbf{Z}_t, t, c, \boldsymbol{\psi}_0)(\sigma_{t+\Delta t} - \sigma_t), \ \textbf{Inv: } \mathbf{Z}_{t-\Delta t} = \mathbf{Z}_t - v_\theta(\mathbf{Z}_t, t, c, \boldsymbol{\psi}_1)(\sigma_t - \sigma_{t-\Delta t}), \tag{7}$$

where timestep t increases from 0 to 1 in the forward process and decreases from 1 to 0 in the reverse process, advancing by Δt at each step. Evidently, the incomplete latent variable $\boldsymbol{\psi}_1$ is available in the reverse process, while the desired latent $\boldsymbol{\psi}_0$ remains unknown at each step. Consequently, the velocity field $v_\theta(\mathbf{Z}_t, t, c, \boldsymbol{\psi}_0)$ cannot be directly evaluated, making the exact generation intractable. In this situation, we introduce a training-free CLA sampling strategy, which progressively enriches incomplete visual conditions using generated multimodal information and iteratively corrects the conditional flow trajectory. Specifically, at the t-th alignment cycle, the model predicts the velocity field $v_t = v_\theta(\mathbf{Z}_t, t, c, \boldsymbol{\psi}_t)$, and numerically integrates from current state to yield the one-step

Algorithm 1: Cyclic Latent Alignment

Input: $\mathbf{X}$, $\mathbf{M}$, c, $\mathcal{E}_\theta$, $\mathcal{D}_\theta$,v_θ, $\{\lambda_t\}$, $\{\sigma_t\}$.

$\mathbf{Z}_1 \sim \mathcal{N}(0, \mathbf{I})$, $\boldsymbol{\psi}_1 = \mathcal{E}_\theta(\mathbf{M} \odot \mathbf{X})$

for $t = 1, 1 - \Delta t, \ldots, \Delta t$ **do**

$v_t \leftarrow v_\theta(\mathbf{Z}_t, t, c, \boldsymbol{\psi}_t)$

$\widehat{\mathbf{Z}}_0^t \leftarrow \mathbf{Z}_t - (\sigma_t - \sigma_0)v_t$

$\boldsymbol{\psi}^* = \mathbf{M} \odot \mathbf{X} + (1 - \mathbf{M}) \odot \mathcal{D}_\theta(\widehat{\mathbf{Z}}_0^t)$

$\boldsymbol{\psi}_{t-\Delta t} = \lambda_t \boldsymbol{\psi}_1 + (1 - \lambda_t)\mathcal{E}_\theta(\boldsymbol{\psi}^*)$

$\mathbf{Z}_{t-\Delta t} \leftarrow \mathbf{Z}_t - (\sigma_t - \sigma_{t-\Delta t})v_\theta(\mathbf{Z}_t, t, c, \boldsymbol{\psi}_{t-\Delta t})$

$\mathbf{X} \leftarrow \mathcal{D}_\theta(\mathbf{Z}_0)$

return $\mathbf{X}$

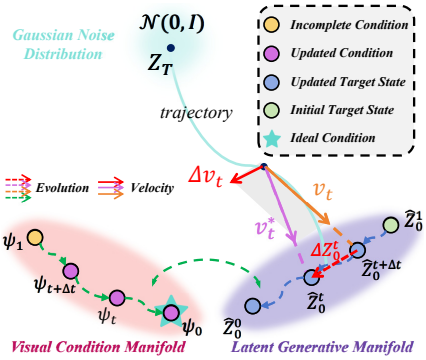

Figure 3: Cyclic Latent Alignment.

generated endpoint state $\widehat{\mathbf{Z}}_0^t$. Instead of recycling the estimate as the visual condition, we decode it and fill only missing entries, preserving observed measurements. Then the enriched observation is re-encoded and interpolated with the initial condition $\boldsymbol{\psi}_1$:

$$\boldsymbol{\psi}_{t-\Delta t} = \lambda_t \boldsymbol{\psi}_1 + (1 - \lambda_t)\mathcal{E}_\theta(\boldsymbol{\psi}^*) \leftarrow \boldsymbol{\psi}^* = \mathbf{M} \odot \mathbf{X} + (1 - \mathbf{M}) \odot \mathcal{D}_\theta(\widehat{\mathbf{Z}}_0^t), \tag{8}$$

where λ_t is the annealed coefficient that decreases from 1 to 0, and $\boldsymbol{\psi}_{t-\Delta t}$ is the updated visual condition. This schedule exploits the generated estimate to compensate for missing modalities, while progressively re-anchoring the condition to the generated evidence as sampling approaches the data manifold. Such interpolation is motivated by the observed proximity of latent representations under different band configurations (Fig. 4) and the observation that richer visual conditions generally improve translation fidelity (Appendix D.3). Then the reverse trajectory is advanced by:

$$\mathbf{Z}_{t-\Delta t} = \hat{\mathbf{Z}}_{t-\Delta t} - (\sigma_t - \sigma_{t-\Delta t})(v_t^* - v_t) \xleftarrow{v_t^* = v_\theta(\mathbf{Z}_t, t, c, \boldsymbol{\psi}_{t-\Delta t})} \hat{\mathbf{Z}}_{t-\Delta t} = \mathbf{Z}_t - (\sigma_t - \sigma_{t-\Delta t})v_t, \tag{9}$$

By repeated alignment, CLA continuously refines the visual condition and steers the transport direction, yielding more reliable velocity estimates without additional training. Each step requires two velocity evaluations, one for endpoint estimation and one for trajectory advancement under the updated condition. The total number of neural function evaluations (NFE) is $2/\Delta t$, scaling linearly with the sampling steps. The detailed algorithm is summarized in Alg. 1 and presented in Fig. 3.

4 Experiments

Experimental Setup. RSI-MM consists of 16 public datasets, from each of which we select a subset of samples for task evaluation (See Appendix C.1). For image reconstruction, we compare RS-VAE against Any2any Chen et al. (2026) with five modality-specific VAEs using PSNR and SSIM Wang et al. (2004); For text-to-image generation, due to the lack of unified methods and metrics across remote sensing modalities, we compare OmniRSI with representative methods including Txt2Img-MHN Xu et al. (2023b), CC-Diff Zhang et al. (2024a), DiffusionSat Khanna et al. (2024), SD3.5 Medium Esser et al. (2024), MMDiff Tang et al. (2026), HSIGene Pang et al. (2026a), Text2Earth Liu et al. (2025), all using their official pretrained weights. Generation quality is evaluated by FID Heusel et al. (2017), NIQE Mittal et al. (2012), CLIPIQA Wang et al. (2023), and CLIP-Score Hessel et al. (2021) on RGB bands; For multimodal translation, we report PSNR and SSIM and compare OmniRSI with DDPM Song et al. (2021), ReFlow Liu et al. (2023), DDBM Zhou et al. (2024), RDDM Liu et al. (2024a), and RDBM Wang et al. (2026a). These baselines are adapted by modifying their input/output dimensions to match slot number and retrained separately for each task.

Implementation Details. OmniRSI is implemented in PyTorch and trained in three stages. Firstly, RS-VAE is optimized using $\mathcal{L}_{vae}$ with empirically selected weights $\alpha = 0.1$, $\beta = 0.00025$, $\gamma = 0.01$. Secondly, SD 3.5 Medium is initialized from pretrained weights and its MMDiT blocks ($N_b = 24$) are fully fine-tuned, while RS-VAE and text encoders (*i.e., CLIP-G/14, CLIP-L/14, and T5-XXL*) remain frozen. Thirdly, the ControlNet branch with $N_c = 12$ Transformer blocks is initialized from SD3.5 backbone weights and trained with all other components frozen. The coefficient schedules are $\sigma_t = \lambda_t = t$. All training uses AdamW with a learning rate of 3×10^{-5} under cosine decay and proper batch size. Inference uses 40 steps and default seeds, with $\omega = 5.0$ for generation and 1.0 for translation. The total training takes 60 days on 8 NVIDIA A100 80GB GPUs (See Appendix C.2).

Table 2: Quantitative results of our RS-VAE and modality-specific VAEs in Any2any. FLOPS are measured at 256×256 resolution. **Bold** ranks the best and underline ranks the second best.

Method	PAN		OPT		SAR		NIR		MSI		Average		Complexity	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	Params(M)	FLOPs(G)
Any2any	37.68	0.838	39.06	0.951	26.85	0.846	29.99	0.951	21.78	0.796	32.69	0.890	13.322∼83.610	329.120∼473.663
Ours	**42.79**	**0.972**	**39.71**	**0.958**	**27.15**	**0.878**	**32.83**	**0.953**	**26.91**	**0.869**	**34.95**	**0.935**	83.815	447.742

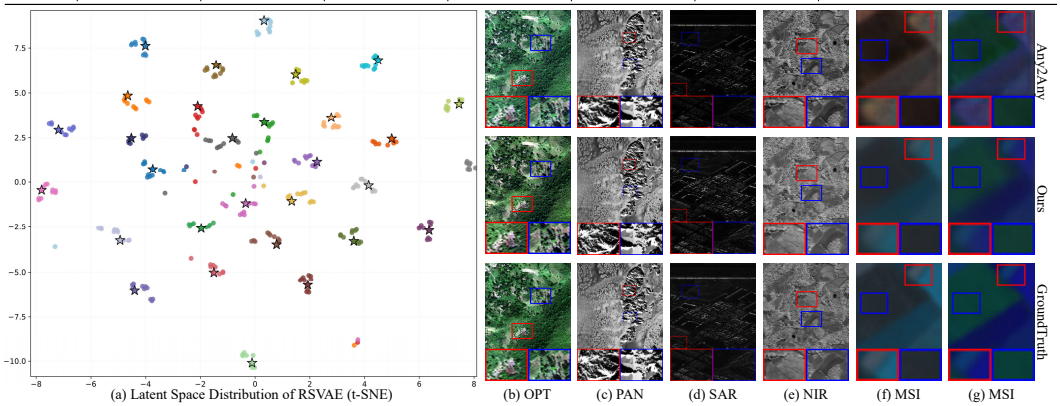

Figure 4: Representation and reconstruction of RS-VAE. (a) Latent distribution, where stars denote different samples and circles are their corresponding subset bands. Most representations cluster tightly around their associated sample, indicating that RS-VAE preserves scene-level consistency well for heterogeneous observations. (b)-(g) Reconstruction comparisons across OPT, PAN, SAR, near-infrared (NIR), and MSI modalities, including two Sentinel-2 composites (B5/B6/B7 and B8A/B11/B12). The top/middle/bottom rows are results of Any2any, RS-VAE, groundtruth (GT).

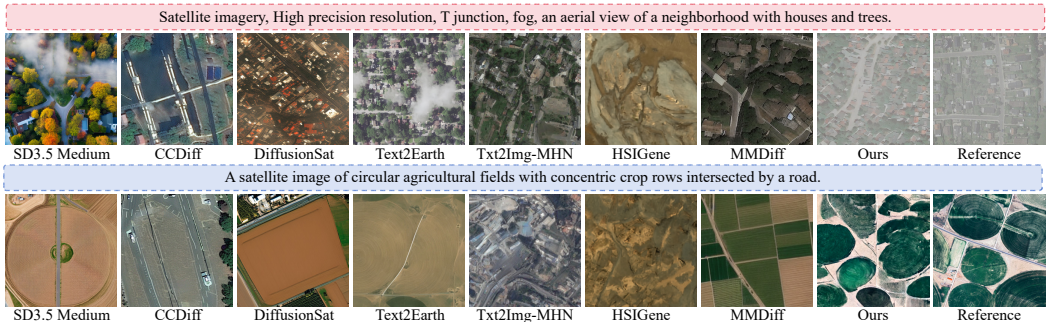

Figure 5: Qualitative results of text-to-image generation on RGB composites.

4.1 Main Results

Image Reconstruction. RS-VAE aims to map heterogeneous observations into a shared latent space. As presented in Fig. 4, full-observation representations overlap with multiple single-band regions, reflecting the integration of complementary modality information. Following task settings of Any2any, RS-VAE consistently achieves the best reconstruction performance across specific modality combinations in Tab. 2. Further band-wise evaluation in Appendix D.1 demonstrates robust self-reconstruction from sparse observations and progressively improved full-band reconstruction with increasing band availability, confirming its effective exploitation of modality dependencies.

Text-to-image generation. For fairness, we use identical text prompts and compare RGB composites extracted from their outputs, since existing methods are typically adapted to modality-specific datasets and output spaces. As illustrated in Fig. 5, OmniRSI produces images with improved semantic alignment, coherent spatial layouts, and faithful scene structure. Quantitative results in Tab. 3 further confirm that OmniRSI achieves the best performance across all four metrics. Moreover, the performance of varied guidance strength in Tab. 4 shows that $\omega = 5.0$ provides the best overall trade-off, while overly weak or strong guidance degrades generation quality. Beyond RGB synthesis, we provide the band-wise visualizations for other modalities in Fig. A2 and Fig. A3 (See Appendix D.2), where the generated data are compared with available ground-truth bands and ex-

Table 3: Results of text-to-image task.

Method	FID Score ↓	NIQE↓	CLIPIQA↑	CLIPScore↑
CC-Diff	132.137	6.168	0.569	0.725
Txt2Img-MHN	150.063	5.398	0.377	0.672
HSIGene	105.162	7.463	0.217	0.662
Text2Earth	52.842	7.362	0.471	0.750
DiffusionSat	81.799	6.912	0.541	0.755
MMDiff	80.922	5.824	0.377	0.780
SD3.5 Medium	93.089	6.397	0.597	0.758
Ours	**28.693**	**4.399**	**0.695**	**0.830**

Table 4: Ablation on varied ω value.

Guidance	FID Score ↓	NIQE↓	CLIPIQA↑	CLIPScore↑
$\omega = 1.0$	42.077	4.790	0.602	0.786
$\omega = 2.0$	35.084	4.767	0.639	0.800
$\omega = 3.0$	30.908	4.793	0.633	0.805
$\omega = 4.0$	29.160	4.864	0.668	0.803
$\omega = 5.0$	**28.693**	**4.399**	**0.695**	**0.830**
$\omega = 6.0$	29.183	4.924	0.661	0.802

Table 5: Quantitative comparisons of translation tasks.

Method	PAN ⟼ MSI		MSI ⟼ PAN		SAR ⟼ MSI		MSI ⟼ SAR		Average		Complexity	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	Params(M)	FLOPs(G)
DDPM	10.88	0.271	12.49	0.335	12.23	0.126	12.81	0.057	12.01	0.232	9.24	34.245
ReFlow	18.72	0.564	29.71	0.813	10.16	0.239	15.46	0.196	20.62	0.533	34.25	9.238
DDBM	16.35	0.020	20.99	0.123	8.45	0.015	9.92	0.032	15.55	0.058	30.09	64.305
RDDM	19.73	0.360	29.92	0.612	12.13	0.253	16.99	0.118	21.59	0.389	60.41	7.495
RDBM	21.02	0.656	31.11	0.881	12.16	0.367	18.01	0.245	22.59	0.616	39.17	7.83
Ours(✗text)	**22.07**	**0.662**	**35.17**	**0.914**	**17.36**	**0.373**	**20.32**	**0.291**	**25.63**	**0.639**	2789.06	1968.955
Ours(✓text)	25.16	0.786	36.10	0.925	18.39	0.382	20.50	0.302	27.10	0.683	2789.06	1968.955

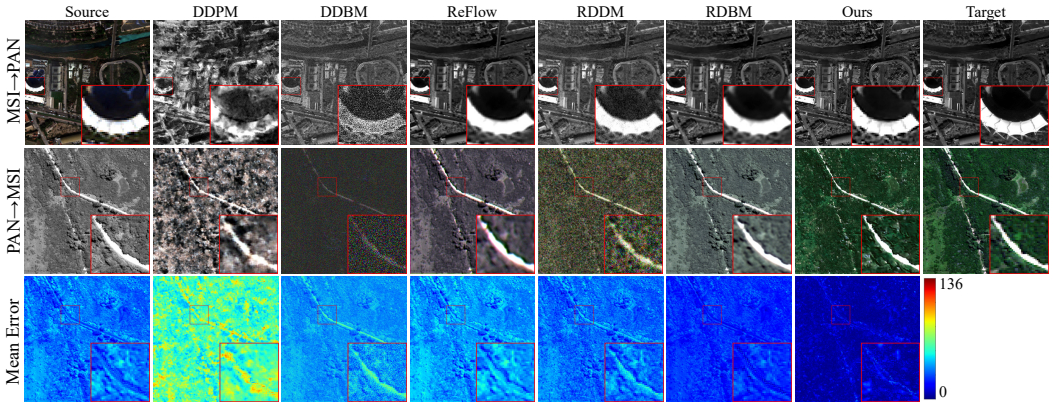

Figure 6: Qualitative results of image-to-image generation. Displayed in RGB composite.

tended bands absent from the reference data. These results indicate that the unified latent space supports coherent multimodal synthesis despite incomplete pairing and partial observations during training, extending OmniRSI beyond specific synthesis toward unified multimodal generation.

Multimodal Image Translation. We assess translation under two representative tasks, PAN↔MSI and SAR↔MSI, where MSI encompasses both RGB and satellite-specific bands, as illustrated in Fig. 6 and Fig. A5. OmniRSI produces the most faithful cross-modal translations, preserving spatial structures while recovering modality-specific characteristics across PAN, MSI, and SAR observations. Quantitative results in Tab. 5 further confirm this advantage. Despite using a unified model for all modality mappings, OmniRSI achieves the highest average PSNR and SSIM across all translation tasks and substantially outperforms other methods separately optimized for individual task, demonstrating the effectiveness of unified latent representation and powerful generation capacity. The effects of textual conditioning and guidance strength ω are reported in Tab. 6. Without text prompts, the conditional and unconditional branches in Eq. 5 share identical text embeddings and visual conditions, causing the guidance residual to vanish and rendering ω ineffective. Textual descriptions provide complementary semantic guidance to visual observations, substantially improving performance, while task-specific guidance strengths further enhance image fidelity. To evaluate the spectral fidelity, we evaluate the generated MSI images via four representative normalized-difference indices: NDVI Tucker (1979), NDWI McFeeters (1996), NDRE Barnes et al. (2000), and NDBI Zha et al. (2003), which characterize vegetation, water, red-edge, and built-up information, respectively. These indices are computed using the corresponding spectral bands available under each sensor configuration (See Appendix D.4), as shown in Tab. 7, OmniRSI achieves the best metrics across all indices, indicating more accurate preservation of inter-band spectral relationships. Fig. 7 further shows that the derived index maps more closely resemble the references, whereas competing methods exhibit noticeable spectral distortions. Since PAN-SAR pairs are unavailable in RSI-MM, we

Table 6: Ablation study on the effect of ω for different translation tasks.

Guidance	PAN ⟼ MSI		MSI ⟼ PAN		SAR ⟼ MSI		MSI ⟼ SAR		Average	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
✕ Text prompts, ✓ Euler, ✓ CLA										
$\omega = 1.0$	22.07	0.662	35.17	0.914	17.36	0.373	20.32	0.291	25.63	0.639
$\omega = 2.0$	22.07	0.662	35.17	0.914	17.36	0.373	20.32	0.291	25.63	0.639
$\omega = 3.0$	22.07	0.662	35.17	0.914	17.36	0.373	20.32	0.291	25.63	0.639
$\omega = 4.0$	22.07	0.662	35.17	0.914	17.36	0.373	20.32	0.291	25.63	0.639
✓ Text prompts, ✓ Euler, ✓ CLA										
$\omega = 1.0$	**25.16**	0.786	**36.10**	**0.925**	**18.39**	0.382	**20.50**	**0.302**	**27.10**	0.683
$\omega = 2.0$	25.08	**0.811**	35.87	0.922	18.25	**0.390**	20.32	0.291	26.94	**0.689**
$\omega = 3.0$	24.27	0.801	35.56	0.918	16.46	0.351	20.30	0.291	26.28	0.678
$\omega = 4.0$	23.35	0.776	35.06	0.912	16.19	0.341	20.22	0.291	25.76	0.666

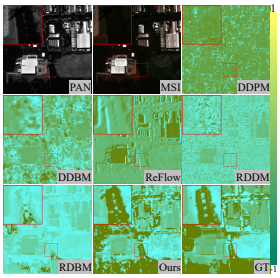

Figure 7: NDVI Results.

Table 7: Quantitative results of NDXI application across diverse translated MSI modalities.

Method	4 Bands (NDVI)		7 Bands (NDWI)		8 Bands (NDRE)		10 Bands (NDBI)		Average	
	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓
DDPM	0.502	0.441	0.341	0.266	0.884	0.871	0.593	0.514	0.648	0.596
ReFlow	0.524	0.484	0.317	0.247	0.837	0.810	0.518	0.479	0.613	0.576
DDBM	0.739	0.586	0.522	0.432	0.790	0.714	0.654	0.557	0.714	0.609
RDDM	0.575	0.466	0.617	0.488	0.765	0.723	0.574	0.509	0.642	0.568
RDBM	0.428	0.373	0.394	0.379	0.422	0.327	0.464	0.393	0.434	0.364
Ours(✕text)	**0.360**	**0.283**	**0.279**	**0.233**	**0.284**	**0.210**	**0.337**	**0.289**	**0.321**	**0.256**
Ours(✓text)	0.255	0.192	0.166	0.129	0.199	0.140	0.147	0.101	0.196	0.142

Table 8: Effect of CLA on different solvers for multimodal image translation tasks.

Method			Step	NFE	PAN ⟼ MSI		MSI ⟼ PAN		SAR ⟼ MSI		MSI ⟼ SAR		Average	
Solver	CLA	Text			PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Euler	✕	✕	40	40	21.73	0.660	34.77	0.910	17.09	0.365	19.54	0.278	25.21	0.632
Euler	✓	✕	40	80	22.07	0.662	35.17	0.914	17.36	0.373	20.32	0.291	25.63	0.639
Euler	✕	✓	40	40	24.30	0.768	35.48	0.919	18.11	0.375	19.66	0.283	26.43	0.671
Euler	✓	✓	40	80	25.16	0.786	36.10	0.925	18.39	0.382	20.50	0.302	27.10	0.683
Euler	✕	✕	80	80	21.78	0.653	35.09	0.907	17.19	0.366	20.27	0.285	25.48	0.631
Euler	✕	✓	80	80	24.58	0.776	35.96	0.925	18.21	0.369	20.39	0.301	26.83	0.678
DPMSolver	✕	✕	40	40	21.53	0.633	34.61	0.911	16.57	0.324	20.13	0.281	25.10	0.619
DPMSolver	✓	✕	40	80	21.73	0.636	34.90	0.913	16.77	0.331	20.20	0.284	25.31	0.622
DPMSolver	✕	✓	40	40	24.30	0.757	35.28	0.921	16.85	0.343	20.34	0.290	26.26	0.664
DPMSolver	✓	✓	40	80	24.31	0.758	35.55	0.922	17.92	0.350	20.41	0.293	26.55	0.667

perform PAN↔SAR translation in a zero-shot manner, as shown in Fig. A4 (See Appendix D.3). These visualizations demonstrate that OmniRSI can transfer across unseen modality pairs without task-specific adaptation, further highlighting its competitive generalization capacities.

Effects of CLA across different solvers. We evaluate CLA with Euler and DPM-Solver++(2M) Lu et al. (2025) using identical random seeds, as shown in Tab. 8. CLA consistently improves the average translation performance of both solvers, with more pronounced gains for Euler, demonstrating the effectiveness of progressively refining incomplete visual conditions during sampling. This difference can be attributed to their numerical characteristics. Euler directly updates the current state using timestep-wise velocity estimates, allowing CLA-induced condition refinement to immediately affect the transport trajectory, whereas DPM-Solver++(2M) incorporates historical model predictions in its multistep updates and is therefore less responsive to timestep-dependent corrections. Textual conditioning further provides complementary semantic guidance and improves translation fidelity across both solvers. Besides, Euler with CLA outperforms Euler with more sampling steps, under the 80 NFE budget, verifying that refining incomplete visual conditions is more effective than simply increasing the sampling resolution. Despite additional model evaluations, Euler with CLA and text achieves the best overall performance, offering a favorable efficiency–accuracy trade-off.

5 Conclusion

We present OmniRSI, a latent diffusion model for unified multimodal remote sensing image generation under heterogeneous and partially observed data. OmniRSI tightly couples a RS-VAE for shared multimodal representation, a fully adapted diffusion backbone for unified generation and translation, and a training-free cyclic latent alignment mechanism for progressively refining incomplete visual conditions during inference. Furthermore, we construct RSI-MM, a large-scale multimodal dataset spanning diverse sensors, resolutions, and modality configurations. Extensive experiments on reconstruction, text-to-image generation, and cross-modal translation demonstrate our superiority.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets (Appendix A) and editing the manuscript to improve readability and language clarity. We have not used generative AI tools for literature searches or discovery, research ideation, drafting manuscript sections, or proving mathematical claims. We take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS AND REPRODUCIBILITY STATEMENT

This work does not involve ethical concerns requiring specific consideration. Core Code is provided in the supplementary materials and will be fully released upon camera-ready submission, together with pretrained weights and datasets.

REFERENCES

Nan Bao, Yifan Zhao, Wenzhuang Wang, and Jia Li. Envisioning beyond the few: Disentangled semantics and primitives for few-shot atypical layout-to-image generation. In *International Conference on Machine Learning*, 2026.

EM Barnes, TR Clarke, SE Richards, PD Colaizzi, JULIO Haberland, M Kostrzewski, Peter Waller, Christopher Choi, E Riley, T Thompson, et al. Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data. In *Proceedings of the International Conference on Precision Agriculture*, volume 1619, 2000.

Romain Beaumont. Large scale openclip: L/14, h/14 and g/14 trained on laion-2b. *LAION. September*, 15:7, 2022.

Haoyang Chen, Jing Zhang, Di Wang, Hebaixu Wang, Shiqin Wang, Pohsun Huang, Jiayuan Li, Haonan Guo, Zheng Wang, and Bo Du. Any2any: Unified arbitrary modality translation for remote sensing. In *International Conference on Machine Learning*, 2026.

Zhenyuan Chen, Chenxi Wang, Ningyu Zhang, and Feng Zhang. Rscc: A large-scale remote sensing change caption dataset for disaster events. *arXiv preprint arXiv:2509.01907*, 2025.

Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IEEE International Geoscience and Remote Sensing Symposium*, pp. 2115–2118. Ieee, 2018.

Solène Debuysère, Nicolas Trouvé, Nathan Letheule, Elise Colin, and Georgia Channing. Sarlo-80: Worldwide slant sar language optic dataset 80cm. *arXiv preprint arXiv:2606.20523*, 2026.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Chen-Chen Fan, Peiyao Guo, Linping Zhang, Kehan Qi, Haolin Huang, Yong-Qiang Mao, Yuxi Suo, Zhizhuo Jiang, Yu Liu, and You He. Smart-ship: A comprehensive synchronized multimodal aligned remote sensing targets dataset and benchmark for berthed ships analysis. *arXiv preprint arXiv:2508.02384*, 2025.

Michael Fuest, Pingchuan Ma, Ming Gui, Johannes Schusterbauer, Vincent Tao Hu, and Björn Ommer. Diffusion models and representation learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Jiayang Gao, Tianyi Zheng, Jiayang Zou, Fengxiang Yang, Shice Liu, Luyao Fan, Zheyu Zhang, Hao Zhang, Jinwei Chen, Peng-Tao Jiang, Bo Li, and Jia Wang. C^2fg: Control classifier-free guidance via score discrepancy analysis. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 34398–34407, June 2026.

Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4630–4651, 2025.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IEEE International Geoscience and Remote Sensing Symposium*, pp. 204–207. IEEE, 2018.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

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

Teng Hu, Jiangning Zhang, Ran Yi, Hongrui Huang, Yabiao Wang, and Lizhuang Ma. Sara: High-efficient diffusion model fine-tuning with progressive sparse low-rank adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 13254–13266, 2025.

Yuqi Hu, Longguang Wang, Xian Liu, Ling-Hao Chen, Yuwei Guo, Yukai Shi, Ce Liu, Anyi Rao, Zeyu Wang, and Hui Xiong. Simulating the real world: A unified survey of multimodal generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *International Conference on Learning Representations*, volume 2024, pp. 5586–5604, 2024.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 11513–11522. IEEE, 2022.

Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106: 102638, 2022.

Jiang Lin, Xinyu Chen, Song Wu, Zhiqiu Zhang, Jizhi Zhang, Ye Wang, Qiang Tang, Qian Wang, Jian Yang, and Zili Yi. Freecontrol: Efficient, training-free structural control via one-step attention extraction. *Advances in Neural Information Processing Systems*, 38:545–576, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Chenyang Liu, Keyan Chen, Rui Zhao, Zhengxia Zou, and Zhenwei Shi. Text2earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. *IEEE Geoscience and Remote Sensing Magazine*, 13(3):238–259, 2025.

Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, and Liangqiong Qu. Residual denoising diffusion models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 2773–2783. IEEE, 2024a.

Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

Wenze Liu, Le Zhuo, Yi Xin, Sheng Xia, Peng Gao, and Xiangyu Yue. Customize your visual autoregressive recipe with set autoregressive modeling. *arXiv preprint arXiv:2410.10511*, 2024b.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Yongxiang Liu, Jie Zhou, Yafei Song, Tianpeng Liu, and Li Liu. Huiyanearth-sar: A foundation model for high-fidelity and low-cost global remote sensing imagery generation. *arXiv preprint arXiv:2604.11444*, 2026.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium*, pp. 3226–3229. IEEE, 2017.

Stuart K McFeeters. The use of the normalized difference water index (ndwi) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7):1425–1432, 1996.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *International Conference on Learning Representations*, volume 2024, pp. 51772–51783, 2024.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Hakki Motorcu and Mujdat Cetin. Pg-controlnet: A physics-guided controlnet for generative spatially varying image deblurring. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 2207–2217, 2026.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4296–4304, 2024.

Li Pang, Xiangyong Cao, Datao Tang, Shuang Xu, Xueru Bai, Feng Zhou, and Deyu Meng. Hsigene: A foundation model for hyperspectral image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):730–746, 2026a.

Yatian Pang, Peng Jin, Shuo Yang, Bin Zhu, Bin Lin, Chaoran Feng, Zhenyu Tang, Liuhan Chen, Francis EH Tay, Ser-Nam Lim, et al. Next patch prediction for autoregressive visual generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8260–8268, 2026b.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 2545–2555. IEEE, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

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

Xinyu Shao, Jianping Zhang, Haozhi Wang, Leo Maxime Brunswic, Kaiwen Zhou, Jiqian Dong, Kaiyang Guo, Zhitang Chen, Jun Wang, Jianye Hao, et al. Generative models in decision making: A survey. *arXiv preprint arXiv:2502.17100*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Datao Tang, Xiangyong Cao, Xuan Wu, Jialin Li, Jing Yao, Xueru Bai, Dongsheng Jiang, Yin Li, and Deyu Meng. Aerogen: Enhancing remote sensing object detection with diffusion-driven data generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 3614–3624. IEEE, 2025.

Haojun Tang, Wenda Zhao, Hengshuai Cui, and Haipeng Wang. Mmdiff: Multi-modal remote sensing image generation via cross-modality spatial feature transfer. *ISPRS Journal of Photogrammetry and Remote Sensing*, 241:93–108, 2026.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322, 2020.

Compton J Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment*, 8(2):127–150, 1979.

Chao Wang, Wei Lu, Xiang Li, Jian Yang, and Lei Luo. M4-sar: A multi-resolution, multi-polarization, multi-scene, multi-source dataset and benchmark for optical-sar object detection. *arXiv preprint arXiv:2505.10931*, 2025a.

Hebaixu Wang, Jing Zhang, Haoyang Chen, Haonan Guo, Di Wang, Jiayi Ma, and Bo Du. Residual diffusion bridge model for image restoration. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 8375–8386, 2026a.

Hebaixu Wang, Jing Zhang, Haonan Guo, Di Wang, Jiayi Ma, Bo Du, and Liangpei Zhang. Universal pansharpening foundation model. *arXiv preprint arXiv:2603.03831*, 2026b.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

Yuqing Wang, Shuhuai Ren, Zhijie Lin, Yujin Han, Haoyuan Guo, Zhenheng Yang, Difan Zou, Jiashi Feng, and Xihui Liu. Parallelized autoregressive visual generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 12955–12965. IEEE, 2025b.

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

Kang Wu, Yingying Zhang, Lixiang Ru, Bo Dang, Jiangwei Lao, Lei Yu, Junwei Luo, Zifan Zhu, Yue Sun, Jiahao Zhang, et al. A semantic-enhanced multi-modal remote sensing foundation model for earth observation. *Nature Machine Intelligence*, 7(8):1235–1249, 2025.

Junshi Xia, Hongruixuan Chen, Clifford Broni-Bediako, Yimin Wei, Jian Song, and Naoto Yokoya. Openearthmap-sar: A benchmark synthetic aperture radar dataset for global high-resolution land cover mapping [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 13(4): 476–487, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304. IEEE, 2025.

Chenxi Xie, Minghan Li, Shuai Li, Yuhui Wu, Qiaosi Yi, and Lei Zhang. Dnaedit: Direct noise alignment for text-guided rectified flow editing. *Advances in Neural Information Processing Systems*, 38:124456–124474, 2026.

Fang Xu, Yilei Shi, Patrick Ebel, Wen Yang, and Xiao Xiang Zhu. Multimodal and multiresolution data fusion for high-resolution cloud removal: A novel baseline and benchmark. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2023a.

Yonghao Xu, Weikang Yu, Pedram Ghamisi, Michael Kopp, and Sepp Hochreiter. Txt2img-mhn: Remote sensing image generation from text using modern hopfield networks. *IEEE Transactions on Image Processing*, 32:5737–5750, 2023b.

Ziqi Ye, Shuran Ma, Jie Yang, Xiaoyi Yang, Yi Yang, Ziyang Gong, Xue Yang, and Haipeng Wang. Object fidelity diffusion for remote sensing image generation. In *International Conference on Learning Representations*, volume 2026, pp. 3258–3278, 2026.

Zhiping Yu, Chenyang Liu, Liqin Liu, Zhenwei Shi, and Zhengxia Zou. Metaearth: A generative foundation model for global-scale remote sensing image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1764–1781, 2024.

Zhiping Yu, Chenyang Liu, Jinqi Cao, Qinzhe Yang, Siwei Yu, Zhengxia Zou, and Zhenwei Shi. Metaearth-mm: Unified multimodal remote sensing image generation with scene-centered joint modeling. *arXiv preprint arXiv:2605.20090*, 2026.

Zhenghang Yuan, Zhitong Xiong, Lichao Mou, and Xiao Xiang Zhu. Chatearthnet: A global-scale image-text dataset empowering vision-language geo-foundation models. *Earth System Science Data Discussions*, 2024:1–24, 2024.

Yong Zha, Jay Gao, and Shaoxiang Ni. Use of normalized difference built-up index in automatically mapping urban areas from tm imagery. *International Journal of Remote Sensing*, 24(3):583–594, 2003.

Jinqi Zhang, Linhao Cong, Di Zhuang, Lamei Zhang, and Bin Zou. Fsar-cap: A fine-grained two-stage annotated dataset for sar image captioning. *IEEE Geoscience and Remote Sensing Letters*, 2026a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the International Conference on Computer Vision*, pp. 3813–3824, 2023.

Mu Zhang, Yunfan Liu, Yue Liu, Yuzhong Zhao, and Qixiang Ye. Cc-diff: enhancing contextual coherence in remote sensing image synthesis. *arXiv preprint arXiv:2412.08464*, 2024a.

Mu Zhang, Yunfan Liu, Yue Liu, Yuzhong Zhao, and Qixiang Ye. Cc-diff++: Spatially controllable text-to-image synthesis for remote sensing with enhanced contextual coherence. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Pan Zhang, Baochai Peng, Chaoran Lu, Quanjin Huang, and Dongsheng Liu. Asanet: Asymmetric semantic aligning network for rgb and sar image land cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:574–587, 2024b.

Xu Zhang, Jianzhong Huang, and Lefei Zhang. Any2rsi: Controllable remote sensing text-to-image generation via any control and enriched description. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12852–12860, 2026b.

Zuopeng Zhao, Ying Liu, Xiaoyu Li, Su Luo, Lu Li, and Wenwen Liu. Anyband-diff: A unified remote sensing image generation and band repair framework with spectral priors. *arXiv preprint arXiv:2605.14341*, 2026.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *International Conference on Learning Representations*, volume 2024, pp. 8160–8171, 2024.

A TEXTUAL ANNOTATION PIPELINE

The datasets in Tab. 1 that lack textual annotations generally comprise paired RGB modalities. We exploit the availability of RGB images, either directly provided or extracted from the corresponding bands subsets, to establish a consistent visual basis for annotation. We feed these RGB visualizations to GPT-5.4-mini Singh et al. (2025) using a unified, evidence-oriented prompt, as shown in box A. The prompt covers eight aspects: scene category, primary objects, secondary elements, visual attributes, spatial layout, spatial relations, landuse Inference, and uncertainty. It prioritizes visible evidence and distinguishes inferred functions from direct observations. Each annotation forms a single coherent plain-text paragraph, as exemplified in Fig. A1. Annotation quality is further examined through random sampling and expert review.

A. The detailed textual descriptions of utilzed annotation prompt

Prompt Template You are an expert in remote sensing image interpretation. Analyze the input remote sensing image in a structured and fine-grained manner.
Instructions: Focus on objective visual evidence first, then make cautious semantic inferences. Use professional remote sensing terminology when appropriate. Be as detailed as possible, but do not invent objects or relations that are not visually supported. If some regions are ambiguous due to resolution, shadow, cloud, occlusion, or image quality, explicitly state the uncertainty. Please output the analysis using the following fields:
1. Overall Scene Type: Identify the overall scene category or categories, such as urban area, industrial zone, port, airport, farmland, forest, river, coastal area, desert, mountain, or others.
2. Major Objects and Elements: List the main visible objects and land-use elements, such as buildings, roads, bridges, vehicles, ships, containers, runways, etc.
3. Minor or Secondary Elements: Describe smaller-scale or less dominant objects and scene elements that are still visually meaningful.
4. Visual Attributes: Describe the visible attributes of each object or region, including shape, size, color, texture, density, regularity, boundary clarity, orientation, etc.
5. Spatial Layout and Distribution: Describe how objects are distributed across the image. Consider patterns such as clustered, scattered, grid-like, linear, radial or irregular. Mention relative positions (center, edge, upper/lower/left/right regions if useful).
6. Spatial and Topological Relationships: Describe relations among objects, such as adjacent to, connected to, crossing, parallel to, surrounding, enclosed within, aligned along, separated by, bordering, intersecting, or extending toward.
7. Functional or Land-Use Interpretation: Based on visible evidence, infer the likely function of the scene or subregions, such as residential, commercial, industrial, logistics, transportation, agricultural, storage, recreational, or under construction.
8. Uncertainty: Explicitly identify objects, regions, or functional interpretations that remain uncertain, ambiguous, or tentative.
Output requirements: Provide a single coherent paragraph in plain text only, not bullet points, within 256 words. Ensure the description is comprehensive yet concise, avoid generic statements, and use natural sentences only.

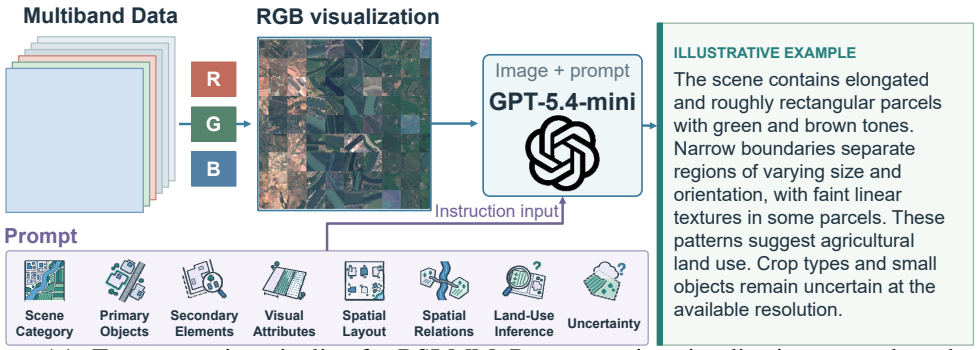

Figure A1: Text annotation pipeline for RSI-MM. Representative visualizations are selected as inputs, followed by GPT-based caption generation with a fixed prompt template.

B Details of Modality-aware Slots

RSI-MM integrates heterogeneous observations from multiple satellite families, including Sentinel, Landsat, WorldView, GaoFen, and GeoEye, whose spectral configurations and modality combinations vary substantially across sensors. To organize heterogeneous remote sensing observations into a consistent input-output structure, we define $K = 18$ modality-aware slots, which comprise one panchromatic slot, sixteen spectral slots spanning coastal to thermal infrared bands, and one SAR slot. Sensor-specific optical bands are mapped to the corresponding canonical slots according to their spectral characteristics, while PAN and SAR observations occupy dedicated slots. This design converts variable sensor configurations into a fixed channel ordering, enabling RS-VAE to jointly process heterogeneous and partially observed data. For each sample, slot availability is indicated by a binary mask $\mathbf{M} = \{\mathbf{m}_k\}_{k=1}^{K}$, where $\mathbf{m}_k = 1$ denotes an observed slot and $\mathbf{m}_k = 0$ denotes a missing one. Notably, these slots serve as ***canonical band references*** rather than exact sensor specifications. The wavelength ranges and centers in Tab. A1 therefore define representative anchors for cross-sensor organization, and bands assigned to the same slot may still differ in spectral response, radiometric characteristics, and acquisition conditions. Even observations from the same sensor and nominal band can vary substantially across seasons or imaging conditions, as exemplified by infrared imagery from the same sensor acquired in different seasons. Thus, the slot design establishes a common representation convention without assuming strict spectral or radiometric equivalence across sensors. Notably, we exclude hyperspectral imagery from the canonical slot space. Its densely sampled narrow bands would require substantially finer slot partitioning and greatly expand the unified input-output space. Moreover, large-scale hyperspectral data are less compatible with the current setting in terms of spatial resolution and scalable semantic annotation, making balanced joint learning more challenging. Extending the canonical space to hyperspectral observations remains an important direction for future work.

Table A1: Canonical slots and their sensor-specific band mappings. Wavelength ranges and centers are slot reference values. Bands assigned to the same slot need not have identical spectral responses.

Slot	Name	Wavelength Range (nm)	Slot Wavelength Center (nm)	Sentinel-2A/B	Landsat 8/9	WorldView2/3	Landsat 7	GaoFen/GeoEye Series
0	PAN	Sensor-dependent	—	—	B8	PAN	B8	PAN
1	Coastal	400–454	427	B1	B1	Coastal	—	—
2	Blue	450–526	488	B2	B2	Blue	B1	Blue
3	Green	510–600	555	B3	B3	Green	B2	Green
4	Yellow	585–625	605	—	—	Yellow	—	—
5	Red	630–690	660	B4	B4	Red	B3	Red
6	RedEdge-1	695—712	703.5	B5	—	—	—	—
7	RedEdge-2	705—748	726.5	B6	—	Red Edge	—	—
8	RedEdge-3	769–793	781	B7	—	—	—	—
9	NIR-1	770–920	845	B8	—	NIR1	B4	NIR1
10	NIR-2	850—1040	945	B8A	B5	NIR2	—	—
11	Water Vapor	932–956	944	B9	—	—	—	—
12	Cirrus	1358–1392	1375	B10	B9	—	—	—
13	SWIR-1	1550–1750	1650	B11	B6	—	B5	—
14	SWIR-2	2090–2350	2220	B12	B7	—	B7	—
15	TIR-1	10600–11190	10895	—	B10	—	—	—
16	TIR-2	11500–12510	12005	—	B11	—	—	—
17	SAR	—	—	—	—	—	—	—

Notes. Sentinel-2A/B band limits are approximated by the reported central wavelength plus or minus half the bandwidth; other bands use the adopted nominal ranges. The GaoFen mappings refer to GF-1/2/7 optical imaging cameras. Unused bands are omitted from this table.

C Datasets Split and Model Details

C.1 Datasets Details.

RSI-MM integrates 16 public datasets to support multimodal image reconstruction, text-to-image generation, and image-to-image translation through the training of the VAE, SD3.5 Medium backbone, and ControlNet branch, respectively. All constituent datasets contribute their training samples to VAE optimization. Samples with image–text pairs, using either existing or newly supplemented captions, are used for text-to-image generation. Image-to-image translation is trained on samples with paired observations across the supported source–target modality configurations. Accordingly, FSARCap, Git-10M, Inria, RSCC, and RSICD are excluded from translation training because they contain only a single image modality.

For comprehensive evaluation, we first use source-provided metadata, including geographic coordinates and cropping information, to identify samples originating from the same larger scene and randomly retain one representative sample per group, thereby reducing scene redundancy. Furthermore, we exclude samples with registration-induced black borders or spatially homogeneous content, such as patches containing only open water or sand. Beyond the redundancy and quality filtering described above, we retain all remaining samples without dataset-level or additional scene-based selection, preserving scene diversity. According to the aforementioned protocol, we initially select 500 samples from each dataset for testing, except for OSCD and PSBench. OSCD is used exclusively for evaluation, whereas PSBench contributes 500 samples from each of its four spectral configurations, yielding 2,000 test samples. All remaining samples are used for training, with consistent partitions maintained across training stages. After filtering, the VAE is evaluated on 3,000 PAN/OPT/NIR samples, 3,500 SAR samples, and 500 MSI samples, following the spectral configuration of Any2Any for multimodal image reconstruction. The SD3.5 Medium backbone is evaluated on 7,000 image samples with associated textual descriptions for text-to-image generation, while the ControlNet branch is evaluated on 4,403 samples with paired observations for image-to-image translation. These counts refer to held-out image samples within each task, without separately counting multiple captions or translation directions associated with the same sample.

Since the source datasets vary substantially in sample numbers and sizes, we adopt dataset-balanced sampling to prevent larger datasets from dominating optimization. For text-to-image training, the text condition is randomly dropped with a probability of 10% to support classifier-free guidance. For ControlNet training, subsets of visual-condition channels are further randomly masked to improve robustness to varying and incomplete modality combinations.

C.2 Model Details.

RS-VAE aims to learn compact latent representations using an encoder $\mathcal{E}_\theta$ and a decoder $\mathcal{D}_\theta$. The encoder parameterizes a diagonal Gaussian posterior $q_\theta(\mathbf{Z} \mid \mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$, where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are the input-dependent posterior mean and variance. Latents are sampled via reparameterization:

$$\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{C.1}$$

where the latent variable $\mathbf{Z}$ is then reconstructed as $\widehat{\mathbf{X}} = \mathcal{D}_\theta(\mathbf{Z})$. This process is optimized based on the evidence lower bound (ELBO):

$$\log p_\theta(\mathbf{X}) \geq \mathbb{E}_{q_\theta(\mathbf{Z}|\mathbf{X})} \left[\log p_\theta(\mathbf{X} \mid \mathbf{Z})\right] - D_{\mathrm{KL}}(q_\theta(\mathbf{Z} \mid \mathbf{X}) \,\|\, \mathcal{N}(0, \mathbf{I})) \,. \tag{C.2}$$

The likelihood term promotes faithful reconstruction through $\mathcal{D}_\theta$, while the KL term regularizes the posterior toward an isotropic Gaussian prior. RS-VAE augments this formulation with the masked reconstruction and latent smoothness constraints, as described in Sec. 3.

After training RS-VAE, we freeze its parameters and encode the entire training set to estimate the global latent mean and variance across samples, channels, and spatial positions. We set the shift factor b to the empirical mean and the scale factor s to the reciprocal of the empirical standard deviation. These global statistics are distinct from the input-dependent posterior parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$. During diffusion training and inference, encoder-produced latents are shifted and scaled before entering the diffusion model. The transformation is reversed before decoding:

$$\widetilde{\mathbf{Z}} = s(\mathbf{Z} - b), \qquad \widehat{\mathbf{X}} = \mathcal{D}_\theta\Big(\widehat{\mathbf{Z}}/s + b\Big) \,, \tag{C.3}$$

where $\mathbf{Z}$ is obtained from $\mathcal{E}_\theta$ and $\widehat{\mathbf{Z}}$ denotes the diffusion output in the normalized latent space. The estimated factors for RS-VAE are $s = 0.8368$ and $b = -0.0881$, replacing the original SD3.5 VAE values of $s = 1.5305$ and $b = 0.0609$. These factors remain fixed throughout diffusion training and inference, standardizing the training latents to approximately zero global mean and unit global variance without requiring an exactly Gaussian distribution.

SD3.5 Medium Backbone comprises 24 multimodal diffusion transformer (MMDiT) blocks. Given a noisy latent $\mathbf{Z}_t \in \mathbb{R}^{16\times H\times W}$, it is partitioned into non-overlapping 2×2 patches through a convolutional projection, producing image tokens with $1,536$ dimensions. Spatial positional embeddings are center-cropped from a fixed 384×384 embedding grid, whose capacity is $147,456$ tokens; the actual token count depends on the input latent resolution. Text tokens are linearly projected from $4,096$ to $1,536$ dimensions. A 256-dimensional sinusoidal timestep embedding and a

Table A2: Architecture configuration of the SD3.5 Medium backbone and ControlNet branch.

Configuration	Value
SD3.5 Medium backbone	
MMDiT blocks	24
Latent input/output channels	16 / 16
Latent patch size	2×2
Hidden dimension	1,536
Attention heads / head dimension	24 / 64
Feed-forward expansion ratio	4
Feed-forward hidden dimension	6,144
Feed-forward activation	GELU (Tanh approximation)
Query/key normalization	RMSNorm, $\epsilon = 10^{-6}$
Block normalization	Non-affine LayerNorm, $\epsilon = 10^{-6}$
Additional image self-attention	Blocks 0–12
Text-token projection	$4{,}096 \to 1{,}536$
Pooled-text embedding MLP	$2{,}048 \to 1{,}536 \to 1{,}536$
Timestep embedding MLP	$256 \to 1{,}536 \to 1{,}536$
Conditioning MLP activation	SiLU
Positional embedding grid	384×384
Positional embedding capacity	147,456 positions
Output projection per token	$1{,}536 \to 64$
ControlNet branch	
Single-stream Transformer modules	12
Noisy/condition latent channels	16 / 16
Latent patch size	2×2
Input fusion	Sum of separate patch embeddings
Hidden dimension	1,536
Attention heads / head dimension	24 / 64
Feed-forward hidden dimension	6,144
Feed-forward activation	GELU (tanh approximation)
Block normalization	Non-affine LayerNorm, $\epsilon = 10^{-6}$
Query/key normalization	None
Explicit positional embeddings	None
Joint image–text attention	None
Pooled conditioning input dimension	1,536
Pooled-text embedding MLP	$1{,}536 \to 1{,}536 \to 1{,}536$
Timestep embedding MLP	$256 \to 1{,}536 \to 1{,}536$
Conditioning MLP activation	SiLU
Timestep input	Noise level $\times\, 1{,}000$
Residual projections	12 separate $1{,}536 \to 1{,}536$ layers
Control strength	1
Residual injection	After every backbone block
Control-feature index at block i	$\lfloor i/2 \rfloor$, $i = 0, \ldots, 23$

$2,048$-dimensional pooled text representation are processed by separate two-layer MLPs with SiLU activation and summed to provide a $1,536$-dimensional modulation signal.

Each MMDiT block maintains separate image and text parameters and performs joint attention over the concatenated image and text queries, keys, and values. Attention uses 24 heads with 64 dimensions per head and RMS normalization of queries and keys. The first 13 blocks additionally perform image-only self-attention. Adaptive LayerNorm and gated residual connections incorporate timestep and pooled-text conditioning, while feed-forward networks use a hidden dimension of 6,144 and GELU activation. The final block uses the text stream to construct attention inputs with-

out further updating its representation. An adaptively normalized linear projection maps each image token to $2\times2\times16$ values, which are rearranged into a latent velocity prediction with the same spatial resolution and channel count as the input. Tab. A2 summarizes the architecture configuration.

ControlNet Branch contains 12 independently parameterized single-stream Transformer modules, which share a block structure similar to MMDiT but differ in their attention interaction manner. The noisy latent and the observed-condition latent, both with 16 channels, are processed by separate 2×2 patch projections and summed to form a shared $1,536$-dimensional token sequence. Each module uses 24-head self-attention with a head dimension of 64, adaptive LayerNorm, gated residual connections, and a feed-forward network with GELU activation. The control branch does not use explicit positional embeddings, query/key normalization, or joint attention with text tokens.

Control conditioning combines a timestep embedding, computed from the noise level multiplied by a certain factor, with a pooled-text embedding. The original $2,048$-dimensional pooled text representation is first mapped to $1,536$ dimensions by the backbone's text embedder and then processed by a separate MLP in ControlNet. Both conditioning MLPs use SiLU activation. Each control-module output is transformed by a dedicated linear projection, producing 12 residual features. With the control strength fixed to 1, these features are injected into the backbone image stream as:

$$\mathbf{H}_i \leftarrow \mathbf{H}_i + \mathbf{h}^c_{\lfloor i/2 \rfloor}, \qquad i = 0, \ldots, 23, \tag{C.4}$$

where $\mathbf{H}_i$ denotes the image-token representation after the i-th backbone block and $\mathbf{h}^c_j$ is the j-th projected control feature. Thus, each control feature is added after each of two consecutive backbone blocks. Notably, the official ControlNet architecture and parameters corresponding to SD3.5 Medium have not been open-sourced. The aforementioned process was modified based on the ControlNet branch of SD3.5 Large; however, the SD3.5 Large model is difficult for us to train.

D More Experiment Results

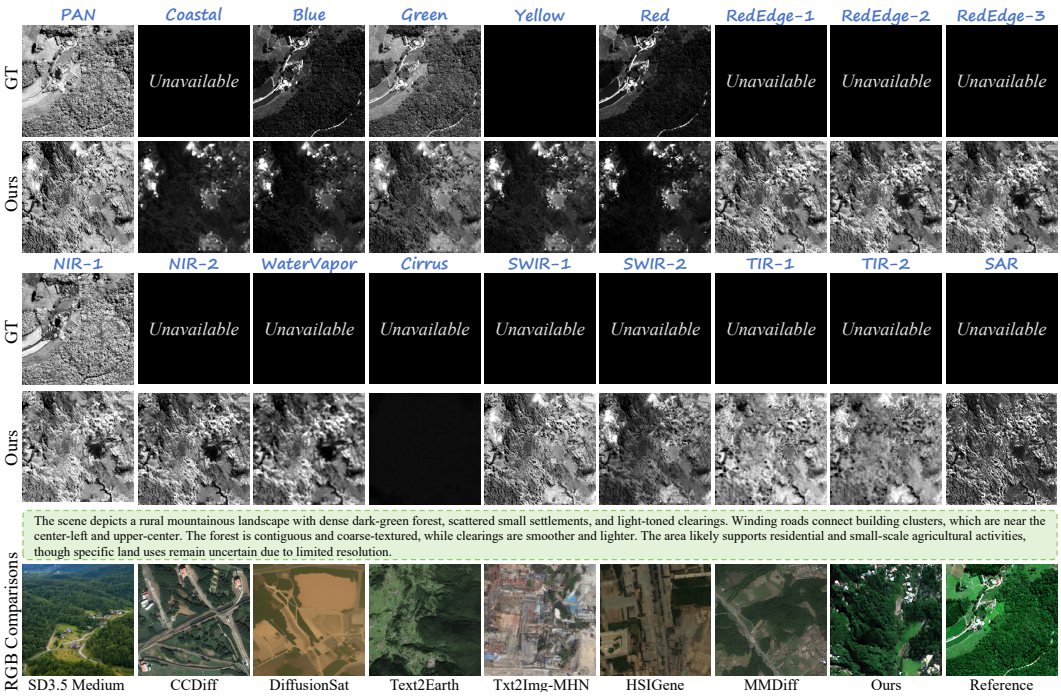

Figure A2: Qualitative results of text-to-image Generation.

D.1 Multimodal Reconstruction

We evaluate reconstruction under varying band availability on a selected subset of the EuroSAT test set, whose Sentinel-2 imagery contains 13 spectral bands. We progressively mask valid bands in a

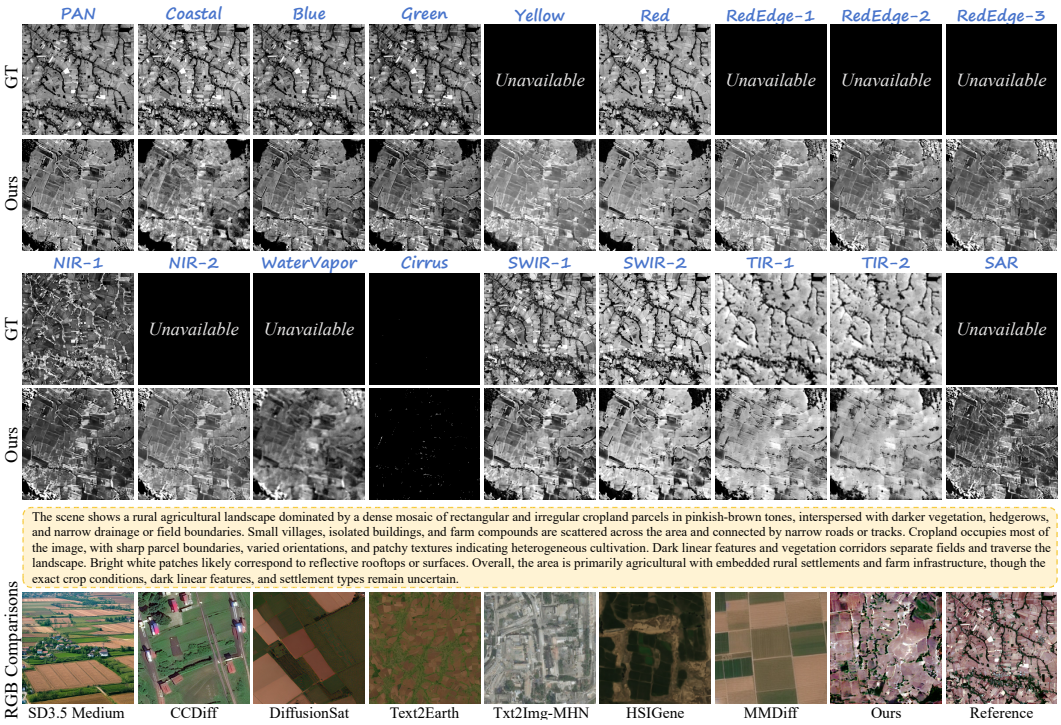

Figure A3: Qualitative results of text-to-image Generation.

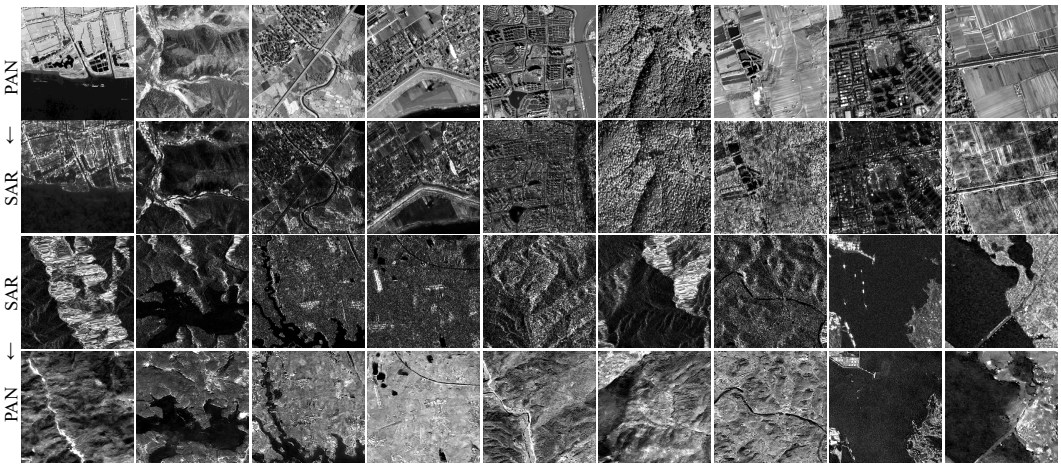

Figure A4: Zero-shot translation from PAN to SAR (top) and from SAR to PAN (bottom).

fixed order and evaluate self-reconstruction (Self-Rec) over the remaining observed bands and full reconstruction (Full-Rec) over all original bands, including those withheld from the input. As shown in Tab. A3, RS-VAE maintains good reconstruction quality even with a single observed band and recovers information in withheld bands, suggesting that it captures useful spectral dependencies. As more bands become available, Full-Rec improves, while Self-Rec exhibits an overall upward trend with fluctuations as the evaluated band subset changes. The two settings coincide when all bands are observed. These results illustrate RS-VAE's ability to reconstruct sparse observations and exploit complementary spectral information for full-band recovery on the evaluated subset.

Table A3: Reconstruction performance of RS-VAE with varying numbers of bands on EuroSAT (50 samples). Self-Rec evaluates the remaining observed bands, whereas Full-Rec evaluates all 13 original bands. Higher PSNR (dB) and SSIM indicate better reconstruction.

Valid Band	1	2	3	4	5	6	7	8	9	10	11	12	13
Self-Rec PSNR	32.66	33.45	34.61	35.61	35.86	37.95	38.95	38.97	39.02	41.44	43.01	43.95	45.58
Self-Rec SSIM	0.930	0.940	0.954	0.959	0.966	0.969	0.974	0.977	0.979	0.986	0.988	0.990	0.991
Full-Rec PSNR	29.75	32.59	33.42	33.45	33.58	36.04	36.82	37.03	37.20	41.15	42.68	43.78	45.58
Full-Rec SSIM	0.904	0.929	0.935	0.936	0.936	0.961	0.968	0.972	0.975	0.987	0.989	0.990	0.991

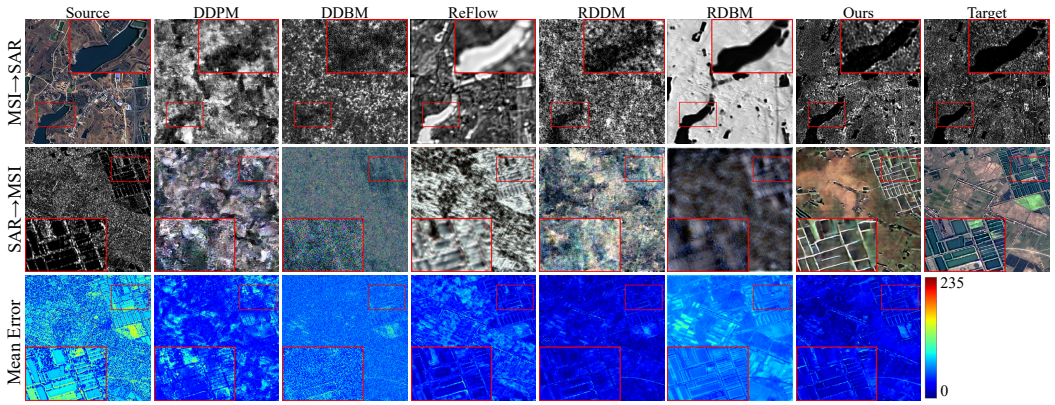

Figure A5: Qualitative results of image-to-image translation. Displayed in RGB.

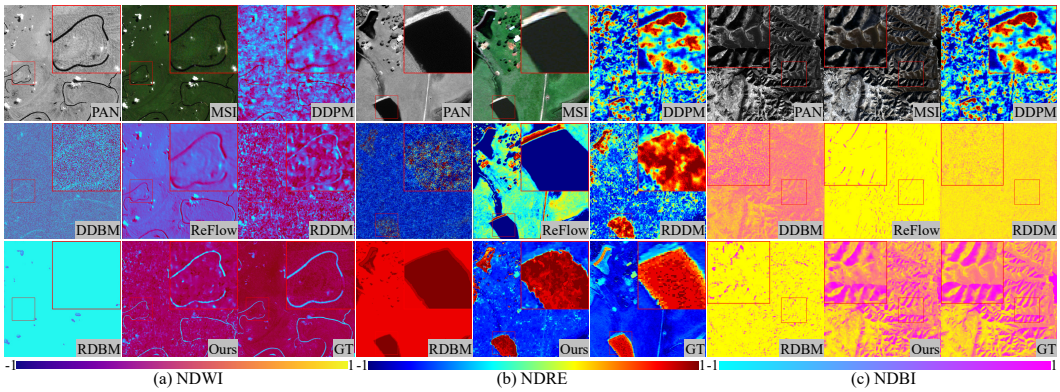

Figure A6: Visualizations of (a) NDWI, (b) NDRE and (c) NDBI (PAN→MSI).

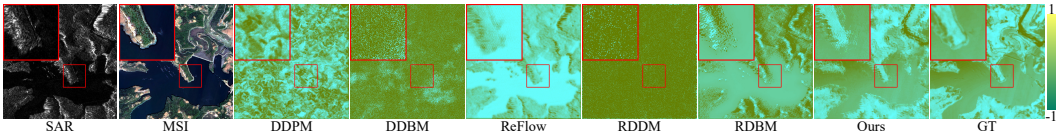

Figure A7: Visualization comparison of NDVI (SAR→MSI).

D.2 Text-to-image Generation

Existing remote sensing generation models predominantly target individual modalities or fixed band configurations. In contrast, OmniRSI models PAN, MSI, and SAR observations within a unified output space. Figs. A2 and A3 complement the RGB comparisons by visualizing all modality slots. Conditioned on a single text prompt, OmniRSI generates spatially consistent structures across heterogeneous spectral and sensing modalities, including those absent from the corresponding reference observation. Black GT panels denote unavailable modalities. For visualization, each band is independently normalized by its maximum value. These results qualitatively demonstrate the capability of OmniRSI for unified full-modality remote sensing generation.

D.3 IMAGE-TO-IMAGE TRANSLATION

OmniRSI supports unified image-to-image translation among three representative remote sensing modality groups, including PAN, MSI, and SAR. The main paper reports quantitative results and representative visualizations for PAN↔MSI translation. Here, we provide additional qualitative comparisons to further examine the cross-modal generation capability of OmniRSI. As shown in Fig. A5, OmniRSI produces visually consistent SAR↔MSI translations while preserving major scene structures and modality-specific appearance characteristics. Since paired PAN–SAR observations are unavailable in RSI-MM, PAN↔SAR translation is evaluated in a zero-shot setting without task-specific paired supervision. As shown in Fig. A4, OmniRSI remains capable of transferring scene structures across previously unpaired modalities and producing plausible target-modality observations. Overall, OmniRSI exhibits consistently strong cross-modal translation performance across both paired and zero-shot modality settings, highlighting its generalization capacity.

To examine the motivation for CLA, we evaluate MSI→PAN translation on PSBench (Landsat9) by progressively masking valid MSI bands in a fixed order. Tab. A4 shows that richer spectral conditions generally improve PAN reconstruction, although performance varies with the retained band subset. Increasing the available bands from one to ten raises both PSNR and SSIM to an extent. These results motivate enriching incomplete visual conditions during sampling, while the solver ablation in the main paper provides direct evidence of CLA's effectiveness.

Table A4: PAN translation performance with varying numbers of observed MSI bands on PSBench (50 samples, MSI→PAN). Bands are progressively masked in a fixed order.

Valid Band	1	2	3	4	5	6	7	8	9	10
PSNR	26.22	30.60	33.32	33.76	34.06	37.36	37.63	39.18	40.92	41.67
SSIM	0.671	0.929	0.937	0.946	0.951	0.965	0.967	0.971	0.974	0.975

D.4 NDXI APPLICATION

To further assess the spectral fidelity of translated MSI observations, we evaluate normalized difference indices (NDXIs) on both PAN→MSI and SAR→MSI tasks. Since the available spectral bands vary across MSI configurations, we compute the corresponding index according to band availability: normalized difference vegetation index (NDVI) Tucker (1979) for four-band MSI, water index (NDWI) McFeeters (1996) for seven-band MSI, red edge index (NDRE) Barnes et al. (2000) for eight-band MSI, and building index (NDBI) Zha et al. (2003) for ten-band MSI. Their definitions are given below:

$$\text{NDVI} = \frac{NIR - R}{NIR + R}, \tag{D.1}$$

$$\text{NDWI} = \frac{G - NIR}{G + NIR}, \tag{D.2}$$

$$\text{NDRE} = \frac{NIR - RE}{NIR + RE}, \tag{D.3}$$

$$\text{NDBI} = \frac{SWIR - NIR}{SWIR + NIR}. \tag{D.4}$$

In accordance with these definitions, the normalized difference indices are confined to the range of $[-1, 1]$, which can serve as the spectral accuracy indicators of the translation process to a certain extent. The quantitative results over these four indices are presented in Tab. 7. Clearly, our method achieves the best values across all these objective metrics, namely root mean squared error (RMSE), and mean absolute error (MAE). Moreover, we present the visualizations in Fig. A6 and Fig. A7. Particularly, all of our index maps exhibit the closest resemblance to the reference, while the other methods suffer from intensity distortions.

D.5 EFFICIENCY COMPARISON

Tab A5 compares peak GPU memory, inference latency, and throughput across resolutions. For image reconstruction, RS-VAE achieves substantially improved computational efficiency over

Any2any, reducing peak memory at both spatial resolutions while attaining higher throughput. This indicates favorable scalability of the proposed VAE for high-resolution remote sensing reconstruction. For text-to-image generation, OmniRSI introduces moderate additional overhead over the SD3.5 Medium backbone owing to the incorporated multimodal conditioning modules. Nevertheless, its peak memory increases only marginally when the resolution is doubled, indicating stable memory scaling. The inference latency increases correspondingly, reflecting the additional conditional processing required by the unified multimodal framework. Multimodal image translation incurs the largest computational cost, as CLA performs additional conditional denoiser evaluations and repeatedly reconstructs and re-encodes intermediate multimodal observations during sampling. Consequently, OmniRSI exhibits higher inference latency at both spatial resolutions. Despite this increased latency, its peak memory remains nearly unchanged across resolutions, suggesting that the major overhead of CLA arises from additional computation rather than memory consumption.

Table A5: Efficiency comparisons among different methods.

Resolution	256×256			512×512		
Method	Mem.(G)	Time(s)	FPS	Mem.(G)	Time(s)	FPS
VAE Reconstruction						
Any2any(MSI)	2.683	0.027	37.037	40.530	0.279	3.579
RS-VAE(Ours)	0.359	0.018	54.969	0.952	0.066	15.110
Text-to-image Generation						
CC-Diff	7.961	2.974	0.336	9.832	5.805	0.172
Txt2Img-MHN	0.581	1.527	0.655	0.980	6.591	0.152
HSIGene	3.896	7.396	0.135	19.044	23.852	0.042
Text2Earth	2.585	1.419	0.705	3.030	1.431	0.699
DiffusionSat	2.612	0.558	1.794	3.051	0.632	1.582
MMDiff	2.172	1.503	0.665	2.617	1.541	0.649
SD3.5 Medium	15.362	1.392	0.719	15.808	1.639	0.610
OmniRSI(Ours)	16.467	2.504	0.399	17.066	3.473	0.288
Multimodal Image Translation						
DDPM	0.384	0.416	2.401	1.295	1.679	0.596
ReFlow	0.367	0.206	4.862	1.226	0.779	1.283
DDBM	0.323	3.024	0.331	0.915	4.706	0.212
RDDM	0.378	0.850	1.176	1.331	3.285	0.304
RDBM	0.637	0.151	6.627	1.689	0.590	1.695
OmniRSI(Ours)	16.482	5.611	0.178	17.111	9.462	0.106

E Datasheets Details

E.1 Motivation

1. *"For what purpose was the dataset created?"*

   A: RSI-MM was created to support unified multimodal remote sensing image generation from heterogeneous and partially observed Earth observation data. Existing resources differ substantially in sensor characteristics, spatial resolution, spectral coverage, modality pairing, and text availability, making it difficult to learn and evaluate shared generative representations. RSI-MM brings together RGB, multispectral (MSI), synthetic aperture radar (SAR), and panchromatic (PAN) imagery with scene descriptions and partially paired observations. It supports multimodal reconstruction, text-to-image generation, and cross-modal translation, and provides the training and evaluation data for OmniRSI.

2. *"Who created the dataset (e.g., which team, research group) and on behalf of which entity?"*

   A: RSI-MM was constructed by the authors of this work. Their name and institutional identities are omitted for anonymous review. The underlying imagery originates from the contributors and providers of the constituent public datasets.

3. *"Who funded the creation of the dataset?"*

   A: The dataset creation was funded by the affiliations of the authors involved in this work.

E.2 COMPOSITION

Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions. Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

1. *"What do the instances that comprise our datasets represent (e.g., documents, photos, people, countries)?"*

   A: Each instance represents a remote sensing image or image patch depicting an Earth surface scene, together with its associated textual descriptions and available modality observations. Depending on the source, an instance may include RGB, multispectral (MSI), synthetic aperture radar (SAR), or panchromatic (PAN) data, or a paired subset of these modalities. RGB composites derived from MSI bands are visual representations of the corresponding observations and do not constitute additional independent acquisitions.

2. *"How many instances are there in total (of each type, if appropriate)?"*

   A: RSI-MM contains 11,518,729 image samples aggregated from 16 public datasets and 11,864,029 text descriptions. The source-wise counts reported in the manuscript sum to 11,518,729: Git-10M (10,452,649), EuroSAT (27,000), OSCD (48), M3M-CR (63,000), OpenEarthMap-SAR (5,033), PIE-RGB-SAR (4,865), WHU-OPT-SAR (4,900), SARLO (119,566), RSCC (62,351), RSICD (10,921), PSBench (449,936), FSARCap (14,480), GID (25,350), Inria (1,800), M4-SAR (112,174), and ChatEarthNet (164,656). These are the listed source counts, not a verified count of globally unique scenes after cross-source deduplication. Details can be found in the Tab 1.

3. *"Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?"*

   A: RSI-MM aggregates 16 public remote sensing datasets and is neither an exhaustive collection nor a uniformly random sample of global Earth observations. Its geographic coverage, scene distribution, spatial resolutions, and modality combinations inherit the selection characteristics of the constituent sources. Existing text annotations are used where available, and descriptions are generated for sources lacking text. The textual annotations are partially created by RSI-MM. Scene redundancy and image-quality filtering are addressed by the protocol in Appendix C.1.

4. *"Is there a label or target associated with each instance?"*

   A:Yes. Textual scene descriptions provide semantic conditioning, while available image bands serve as reconstruction or generation targets. Paired source and target observations support supervised cross-modal translation where such pairs are available. A binary availability mask identifies observed slots; missing modalities are not measured ground truth.

5. *"Is any information missing from individual instances?"*

   A: Yes. Partial modality and band availability is an intentional characteristic of RSI-MM. Observations are organized into 18 canonical slots: one PAN slot, 16 spectral slots spanning coastal to thermal infrared bands, and one SAR slot. A binary mask distinguishes available measurements from absent observations. Not every sample contains all slots, and paired PAN–SAR observations are unavailable in RSI-MM. Missing entries must not be interpreted as physical zero-valued measurements.

6. *"Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?"*

   A: Yes. Relationships are represented through the association of captions with their images and the paired observations supplied by the constituent datasets. The modality-slot mapping identifies how available bands enter the shared representation. However, this organization does not establish pairing between arbitrary images from different sources, guarantee exact temporal correspondence, or imply that bands assigned to the same slot have identical sensor responses.

7. *"Are there recommended data splits (e.g., training, development/validation, testing)?"*

   A: Yes. The manuscript reports randomly selecting partial samples from each constituent dataset for testing. An exact training/validation/test split protocol is provided in Appendix C.1.

8. *"Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?"*

   A: RSI-MM relies on 16 external public datasets for source imagery and, where available, existing captions. Newly generated captions supplement sources without text. It will be released with bundle images, annotations, source metadata, and usage documentation on platforms such as GitHub and Hugging Face, subject to the licenses and terms of the underlying source datasets.

9. *"Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?"*

   A: No. To the best of our knowledge, RSI-MM does not contain confidential records, private communications, or medical, financial, or legal records. The images are derived from publicly released remote sensing datasets, and annotations are constructed for benchmark evaluation.

10. *"Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?"*

    A: No. RSI-MM collection is intended to describe Earth surface scenes and is not designed to include offensive or graphic content.

E.3 Collection Process

In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

1. *"How was the data associated with each instance acquired?"*

   A: The images and available text annotations are obtained from 16 existing public datasets: Git-10M, EuroSAT, OSCD, M3M-CR, OpenEarthMap-SAR, PIE-RGB-SAR, WHU-OPT-SAR, SARLO, RSCC, RSICD, PSBench, FSARCap, GID, Inria, M4-SAR, and ChatEarthNet. For sources without textual annotations, representative visualizations are used to generate scene descriptions. RSI-MM construction is described as dataset aggregation and annotation, without a new image acquisition campaign. Please see Appendix A for details.

2. *"What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?"*

   A: The source datasets provide observations from heterogeneous optical and radar sensors, including Sentinel, Gaofen, PlanetScope, Umbra, and other platforms listed in the manuscript. For missing text annotations, RGB images supplied by the sources or extracted from suitable band subsets were provided to GPT-5.4-mini using a unified, evidence-oriented prompt. The prompt covers scene category, primary objects, secondary elements, visual attributes, spatial layout, spatial relations, land-use inference, and uncertainty. Each generated caption is a single coherent plain-text paragraph. The prompt prioritizes visible evidence and distinguishes inferred functions from direct observations. A systematic human-verification procedure, combining random sampling with expert review, is conducted to ensure the correctness of the annotations.

3. *"If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?"*

   A:RSI-MM is an aggregation of 16 public datasets rather than a sample from a larger set. See Sec. 3 and Appendix C.1 for details.

E.4 Preprocessing, Cleaning, and Labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

1. *"Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?"*
   A: Yes. We extract RGB composites for caption generation, supplement missing textual annotations, and map available observations into modality-aware slots using sensor-specific band mappings. Binary availability masks distinguish missing entries from observed measurements. Optical bands are mapped according to their spectral characteristics, while PAN and SAR occupy dedicated slots. This mapping defines a common channel ordering without establishing radiometric equivalence or spectral harmonization across sensors. The data preparation protocol also includes scene redundancy reduction and filtering of registration-induced black borders and spatially homogeneous samples. See Appendix A for annotation, Appendix C.1 for filtering and splits, and Tab. A1 for band mappings.
2. *"Was the 'raw' data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?"*
   A: Yes. We retained the raw source data in addition to the preprocessed and annotated data. Access to copies distributed by the original source providers remains subject to their availability and terms.
3. *"Is the software that was used to preprocess/clean/label the data available?"*
   A: Yes. The preprocessing and evaluation scripts are planned for public release with the camera-ready version of the paper.

E.5 Uses

The questions in this section are intended to encourage dataset creators to reflect on tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers make informed decisions, thereby avoiding potential risks or harms.

1. *"Has the dataset been used for any tasks already?"*
   A: Yes. RSI-MM is used to train and evaluate OmniRSI for multimodal image reconstruction, text-to-image generation, and multimodal image translation. The study evaluates PAN–MSI and SAR–MSI translation in both directions and presents zero-shot PAN–SAR translation where paired training observations are unavailable. It also assesses spectral fidelity using indices derived from the available bands.
2. *"Is there a repository that links to any or all papers or systems that use the dataset?"*
   A: Yes. we will provide the links on the project GitHub and Hugging Face repositories.
3. *"What (other) tasks could the dataset be used for?"*
   A: RSI-MM may support research on image restoration, multi-source image fusion, image captioning, and multimodal representation learning. It may also support pretraining or data augmentation for downstream detection and segmentation, provided that the task-specific annotations required for training and evaluation are available or separately obtained.
4. *"Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?"*
   A: Yes. Sources differ in spatial resolution, spectral response, acquisition conditions, and modality availability. Captions combine existing annotations with model-generated descriptions based primarily on RGB visualizations; these descriptions may omit non-visible spectral or radar-specific information.
5. *"Are there tasks for which the dataset should not be used?"*
   A: RSI-MM is intended for research on remote sensing representation and generation. Synthetic outputs should not be presented as actual sensor measurements or used as sole evidence for conclusions that require verified observations.

E.6 DISTRIBUTION

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

1. *"Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?"*
   A: The benchmark will be made publicly accessible to the research community, subject to the licenses and terms of the underlying source datasets.
2. *"How will the dataset be distributed (e.g., tarball on website, API, GitHub)?"*
   A: We will provide RSI-MM in the GitHub and the Hugging face repository.
3. *"When will the dataset be distributed?"*
   A: We will create a repository to release the data once the paper is officially published.
4. *"Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?"*
   A: Yes. The newly created RSI-MM instructions, annotations, and metadata will be released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, while source imagery remains subject to the licenses and terms of the original datasets.
5. *"Have any third parties imposed IP-based or other restrictions on the data associated with the instances?"*
   A: Yes. The underlying source images are subject to the licenses and terms of their original datasets. RSI-MM will preserve source attribution and document applicable usage constraints in the release.
6. *"Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?"*
   A: We are not aware of export-control restrictions specific to RSI-MM beyond the licenses and terms of the underlying source datasets. Users are responsible for ensuring compliance with applicable local regulations.

E.7 MAINTENANCE

As with the questions in the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

1. *"Who will be supporting/hosting/maintaining the dataset?"*
   A: The authors of this work will support, host, and maintain the dataset.
2. *"How can the owner/curator/manager of the dataset be contacted (e.g., email address)?"*
   A: They can be contacted via the email addresses listed on the paper or webpage.
3. *"Is there an erratum?"*
   A: There is no erratum for the initial release; updates and known issues will be documented on the project repository.
4. *"Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?"*
   A: Future updates (if any) will be posted on the dataset website.
5. *"Will older versions of the dataset continue to be supported/hosted/maintained?"*
   A:
   Yes. Older versions will be archived when feasible, and updates will be documented on the project repository.
6. *"If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?"*
   A: Yes, we will provide detailed instructions for future extensions.

F Comparison with Closely Related Methods

We discuss MetaEarth-MM, Any2Any, and AnyBand-Diff to clarify their relationship to OmniRSI in terms of modality coverage, representation learning, and generation mechanisms.

Any2Any Chen et al. (2026) addresses unified translation across RGB, SAR, NIR, PAN, and multispectral imagery from Sentinel-2 composites (B5/B6/B7 and B8A/B11/B12). It trains modality-specific VAEs to map heterogeneous observations into latent tensors with a common dimensionality. A shared Diffusion Transformer subsequently learns source-to-target semantic mappings through clean-target latent prediction, conditioned on source and target modality identities. Crucially, dimensional compatibility at the VAE stage does not itself establish cross-modal semantic correspondence: the independent VAE objectives emphasize within-modality reconstruction and prior regularization, without explicitly requiring paired observations to have matching latent semantics. The subsequent translation backbone is responsible for learning these mappings. OmniRSI instead uses a shared slot-based VAE with joint observed-slot and masked reconstruction objectives to learn from heterogeneous, partially observed inputs. The official Any2Any release provides VAE checkpoints, enabling the reconstruction comparison reported in our experiments. However, the translation-backbone checkpoints and inference scripts were not publicly available at the time of checking. Consequently, these VAE checkpoints alone do not permit an end-to-end translation comparison with the official Any2Any model.

MetaEarth-MM Yu et al. (2026) supports joint generation and any-to-any translation across five modalities: RGB, SAR, NIR, PAN, and OpenStreetMap (OSM). It adopts a scene-centered joint flow-matching formulation that separates scene inference from modality-specific generation. A scene inference module extracts intermediate scene representations from noisy modality latents, while a modality-aware routed generator predicts the corresponding velocity fields. OmniRSI instead organizes heterogeneous observations into canonical modality-aware slots and learns a shared RS-VAE for partially observed inputs, followed by text-conditioned generation and visually conditioned translation. Its CLA mechanism further updates the visual condition during sampling by incorporating generated missing entries. These approaches therefore differ in both representation organization and inference-time conditioning. At the time of checking, the official MetaEarth-MM repository contained only a README and license, without implementation code or model checkpoints, preventing direct evaluation using an official release.

AnyBand-Diff ? focuses on optical image generation and spectral recovery, including 3/5/7 bands multispectral reconstruction. It employs dual stochastic masking: sensor-response simulation constructs heterogeneous observations, and random dropout introduces additional missing measurements. Physics-Guided Sampling uses gradients of differentiable physical consistency objectives to refine generation, while a Multi-Scale Physical Loss constrains spectral and spatial statistics at multiple scales. Its emphasis is therefore on spectral reconstruction and physical consistency. OmniRSI addresses generation and translation across PAN, MSI, and SAR within a shared latent framework; its current slot configuration excludes hyperspectral imagery. Moreover, CLA updates the visual conditioning representation through decoded completion and re-encoding, rather than applying gradients from a physical consistency objective. These mechanisms address different aspects of incomplete-observation generation. At the time of checking, we could not identify an official public GitHub repository or downloadable model checkpoints for AnyBand-Diff, and thus could not conduct a direct comparison using an official implementation.

In conclusion, OmniRSI's distinction lies in its combination of a shared slot-based representation, text- and image-conditioned generation, and iterative visual-condition refinement.

G Discussions, Limitations, and Future Work

Limitations and broader impact. OmniRSI aims to unify multimodal remote sensing reconstruction, generation, and translation, but several limitations remain. First, partial modality pairing limits supervision for certain cross-modal mappings. Second, modality-aware slots provide a consistent representation structure but do not explicitly harmonize sensor-specific spectral responses, spatial resolutions, or radiometric characteristics. Consequently, generated observations in missing slots should be interpreted as conditional predictions rather than physically calibrated measurements. Third, captions derived from RGB visualizations may overlook spectral or radar-specific informa-

tion and introduce annotation errors. Last, full backbone fine-tuning requires substantial computational resources, while CLA improves the performance but requires two velocity evaluations per sampling step, increasing inference cost. Despite these limitations, OmniRSI provides a foundation for learning shared generative priors from heterogeneous and incomplete observations. Its ability to synthesize missing modalities may support data augmentation and Earth observation research.

Future Work. Future research could pursue several directions. (1) Expanding and balancing RSI-MM with more geographically diverse, temporally aligned multimodal observations and improving caption quality through modality-aware annotation and verification. (2) Incorporating sensor metadata, spectral response functions, and physical constraints to improve cross-sensor consistency and the reliability of generated spectral information. (3) Developing more efficient architectures and sampling strategies, including distillation of cyclic latent alignment, to reduce training and inference costs and support higher-resolution imagery. (4) Introducing uncertainty estimation for modality combinations with limited or no paired supervision.