Any2RSI: Controllable Remote Sensing Text-to-Image Generation via Any Control and Enriched Description

Xu Zhang[1], **Jianzhong Huang**[1*], **Lefei Zhang**[1]

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
{zhangx0802, hjz_2000, zhanglefei}@whu.edu.cn

Abstract

Recent advances in controllable text-to-image (T2I) generation have achieved impressive results in natural images, but remote sensing (RS) T2I generation remains challenging due to the unique nature of geospatial data. Existing methods struggle to integrate diverse spatial controls and model complex spatial relationships, often failing to maintain semantic consistency with typically vague or incomplete textual descriptions. Moreover, limited by small-scale, low-quality datasets, these models produce outputs with inconsistent layouts and unrealistic content. To address these issues, we propose Any2RSI, a flexible framework for controllable RS T2I generation. It features a Cross-Modal Multi-Control Adapter that extracts modality-agnostic embeddings from heterogeneous spatial inputs, enabling precise spatial guidance. To compensate for sparse or ambiguous text prompts, we introduce a VLM-Empowered Enriched Description Generation module that enhances input descriptions with cross-modal semantics for more coherent T2I generation. Furthermore, we present RST2I-110K, a dataset of over 115,000 high-quality RS image-text pairs across diverse scenes, addressing the current lack of semantically grounded textual annotations needed for RS T2I generation. Any2RSI achieves state-of-the-art performance on both existing and new datasets, improving the realism and structural accuracy of generated RS imagery.

Code — https://github.com/House-yuyu/Any2RSI

Introduction

Remote sensing (RS) text-to-image (T2I) generation synthesizes photorealistic RS images from textual descriptions of geospatial scenes, transforming abstract semantics into visually coherent and spatially accurate imagery. By bridging human language and machine-processable RS data, it enhances the interpretability and accessibility of geospatial information. With growing demand in applications such as environmental monitoring, urban planning, disaster response, and agriculture (Li et al. 2025a, 2024b; Lu et al. 2023; Xu et al. 2025; Li et al. 2025b; Himeur et al. 2022; Lu et al. 2024; Rong et al. 2025; Lu et al. 2025), RS T2I enables intuitive data synthesis and scenario simulation through realistic, semantically meaningful image generation.

*Corresponding author.

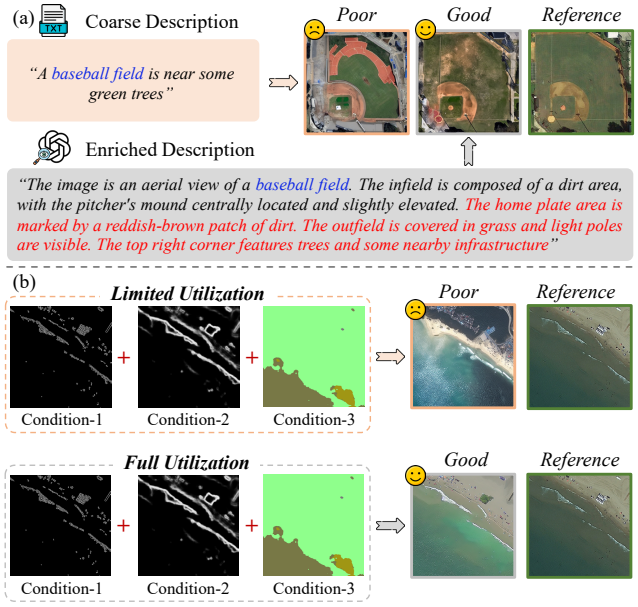

Figure 1: This figure illustrates the difference in the impact on generation performance between using rich descriptions and the ability to fully utilize multiple control conditions during the process of RS image generation.

However, existing RS T2I methods still face two major limitations: **1)** Data level: Current datasets suffer from limited scale, low image quality, monotonous scenes, and coarse textual descriptions, making it difficult to capture complex geospatial patterns. **2)** Method level: Despite advances in generation quality (Xu et al. 2023; Tang et al. 2024), most approaches fail to effectively integrate diverse spatial controls (e.g., edges, segmentation masks). This rigid control hinders modeling of spatial dependencies and synergistic interactions among geographic elements, leading to structurally inconsistent layouts. As shown in Fig. 1(a), models using coarse descriptions (e.g., from RSICD (Lu et al. 2017)) produce semantically inconsistent results, such as generating objects not mentioned in the text, due to over-reliance on spatial constraints without sufficient textual grounding. Moreover, even with detailed descriptions (Fig. 1(b)), methods that poorly fuse multiple control con-

ditions still generate suboptimal outputs. They struggle to model synergies across control conditions and fail to achieve cross-modal alignment between textual and spatial inputs, resulting in comparatively poorer generation quality.

To address these challenges, we propose Any2RSI, a controllable RS T2I generation method that improves control flexibility and semantic consistency. First, to enable precise spatial guidance from heterogeneous control signals, we design the Cross-Modal Multi-Control Adapter (CMMCA). This module extracts modality-agnostic embeddings through two complementary mechanisms: a Cross-Modal Context Aggregation unit that aligns textual semantics with visual queries via self-attention, and a Multi-Control Cross-Attention unit that uses learnable queries to fuse spatial cues from multiple control conditions. Second, recognizing that RS textual prompts are often sparse or ambiguous, we introduce a VLM-Empowered Enriched Description Generation module that enhances input descriptions with cross-modal semantics, yielding richer, more accurate guidance. Third, to alleviate data scarcity, we present RST2I-110K, a large-scale dataset of over 115,000 high-quality image-text pairs across diverse scenes, expanding resources for RS T2I generation research.

In summary, our main contributions are as follows:

- We propose Any2RSI, a controllable RS text-to-image generation framework that supports the flexible combination of multiple spatial control conditions while ensuring semantic consistency between text and image.
- We design a Cross-Modal Multi-Control Adapter that integrates diverse spatial control signals through alternating cross-modal aggregation and attention mechanisms, ensuring semantically coherent image generation.
- We introduce a VLM-Empowered Enriched Description Generation module that leverages detailed ground-object descriptions as additional semantic guidance to further refine the generation results.

Related Work

Text-to-Image Generation

Text-to-Image (T2I) generation, which synthesizes images from textual descriptions, has evolved significantly. Early GAN-based methods (Reed et al. 2016; Zhang et al. 2017) achieved moderate resolution but suffered from training instability. The advent of autoregressive models like DALL-E (Ramesh et al. 2021) enabled strong zero-shot capabilities, while diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021; Rombach et al. 2022) have become dominant due to their superior sample quality and flexible conditioning mechanisms (e.g., cross-attention (Rombach et al. 2022; Zhang et al. 2025a), latent-space modeling (Huang et al. 2025; Zhang et al. 2025b, 2023)).

However, most T2I research focuses on natural images, where semantics are object-centric and spatial layouts are relatively unconstrained. In contrast, RS imagery involves complex geospatial semantics, structured land-cover patterns, and strict geometric relationships. Recent efforts have begun adapting T2I to RS, such as hierarchical prototype learning for semantic alignment (Xu et al. 2023) and diffusion frameworks (Khanna et al. 2024). Yet, these works still rely heavily on text alone, which is often sparse or ambiguous in describing spatial configurations.

Controllable Image Generation

To overcome the limitations of text-only guidance, controllable image generation incorporates auxiliary spatial conditions (Avrahami et al. 2023) to enforce structural fidelity. General-purpose frameworks like ControlNet (Zhang, Rao, and Agrawala 2023) and GLIGEN (Li et al. 2023b) enable flexible conditioning in natural image domains by injecting control signals into diffusion backbones. Inspired by these advances, unified multi-control models (Zhao et al. 2023; Qin et al. 2023; Chen et al. 2025; Xiao et al. 2025; Wang et al. 2025; Sun et al. 2024; Tan et al. 2025; Zhang et al. 2025c; Li et al. 2025c; Pan et al. 2025) have emerged, allowing a single network to handle diverse control types, improving practicality for real-world applications. However, these methods remain largely designed for natural scenes and fail to account for RS-specific challenges, such as the need for physically plausible rendering under environmental constraints.

Recent efforts have begun adapting controllable generation to RS (Tang et al. 2024; Sastry et al. 2024; Goktepe et al. 2025; Pang et al. 2025). For instance, while CRS-Diff (Tang et al. 2024) enables controllable RS image generation by fusing text and spatial guidance in a diffusion model, EcoMapper (Goktepe et al. 2025) generates satellite imagery that reflects climatic conditions by conditioning on environmental variables. However, these approaches lack a unified mechanism to jointly model diverse spatial conditions in a flexible and semantically consistent manner, and they neglect the importance of rich textual semantics.

Remote Sensing Text-to-Image Datasets

In recent years, several datasets (Lu et al. 2017; Zhang et al. 2024; Cheng et al. 2022; Liu et al. 2025) have been introduced to support RS T2I generation. Among them, RSICD (Lu et al. 2017) is widely used but suffers from limited scale, low-resolution imagery, and simplistic textual descriptions with minimal fine-grained annotations. On the other hand, Git-10M (Liu et al. 2025) provides a massive-scale collection, yet its extreme size comes with high scene redundancy. This not only increases computational overhead during training but also dilutes semantic learning due to repetitive or irrelevant samples. To overcome these issues, we introduce RST2I-110K, a large-scale, multi-scene dataset designed specifically for RS T2I generation. It contains over 115,000 high-quality image-text pairs with detailed descriptions covering both global scene attributes and local object-level information.

Methodology

Overview Pipeline

As shown in Fig. 2, Any2RSI comprises three core components. First, the VLM-Empowered Enriched Description Generation (EDG) enhances input text by leveraging

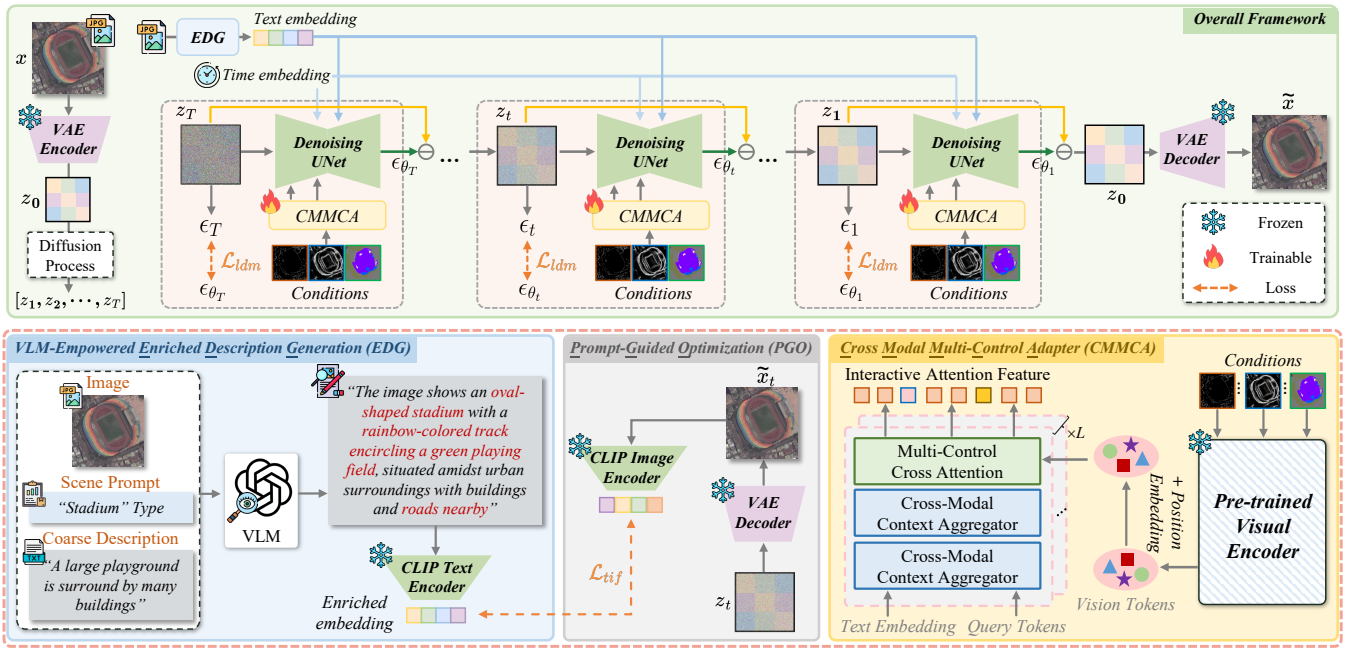

Figure 2: An overview of the proposed Any2RSI architecture. Any2RSI integrates three key components: (1) the EDG to enrich sparse textual prompts with cross-modal semantics; (2) the PGO to refine intermediate outputs via CLIP-based alignment; and (3) the CMMCA to fuse heterogeneous control signals through multi-control attention and cross-modal context aggregation.

cross-modal semantics from a VLM, producing richer and more accurate descriptions to guide semantically coherent image generation. Second, the Prompt-Guided Optimization (PGO) improves text-image semantic alignment during the diffusion process and mitigates early-stage deviations caused by noise accumulation or unstable predictions. Third, the Cross-Modal Multi-Control Adapter (CMMCA) enables fine-grained spatial and semantic control through a multi-level fusion architecture, using enhanced text to better integrate heterogeneous control conditions. Together, these components enable Any2RSI to effectively fuse diverse inputs and generate high-quality, semantically consistent RS images under informative textual guidance.

VLM-Empowered Enriched Description Generation

Manually constructing large-scale RS image-text datasets is costly, time-consuming, and prone to inconsistencies due to subjective interpretations. Moreover, the limited length of human-written descriptions often results in vague or incomplete object annotations, compromising the authenticity and richness of generated images. To address these issues and improve generation quality, we propose EDG, which automatically generates detailed and semantically rich descriptions by leveraging cross-modal understanding and contextual reasoning.

Specifically, we employ InternVL2.5-8B (Chen et al. 2024) to generate detailed and semantically rich descriptions for given scene types. Given an input image $\boldsymbol{x}$, original coarse description $\boldsymbol{T}$, and a scene prompt $\boldsymbol{P}_s$ (e.g., "stadium"), we design a structured prompting strategy that guides the model to identify not only the primary scene but also salient objects, spatial layouts, environmental context, and other visual attributes. This yields fine-grained, context-aware descriptions that go beyond basic labeling, capturing both global and local semantics. For existing datasets like RSICD, which contain $\boldsymbol{T}$, we use these captions as prompts to generate enhanced descriptions via the VLM. In contrast, for our newly constructed RST2I-110K, we directly generate high-quality annotations from scratch using EDG without coarse description, eliminating reliance on low-quality initial labels. The enriched descriptions are then encoded into text embeddings using the CLIP text encoder. The entire process can be described as:

$$c' = VLM(\boldsymbol{P}_s, \boldsymbol{x}, \boldsymbol{T}), \quad \mathcal{T}_e = \Phi_{\mathcal{T}}(c'), \tag{1}$$

where, $\Phi_{\mathcal{T}}(\cdot)$ denotes the CLIP text encoder. c' and $\mathcal{T}_e$ represent the enriched text and text embedding, respectively.

Prompt-Guided Optimization

Even with enriched textual descriptions, generated images may drift from target semantics during diffusion, as most models rely solely on the initial text embedding without enforcing semantic consistency at each denoising step. This causes early errors to accumulate, resulting in misalignment between the output and the intended description.

To enhance semantic alignment and mitigate early-stage deviations, we introduce a CLIP-based generation loss at diffusion time step t, integrating semantic guidance from detailed textual descriptions throughout the generation process. Unlike traditional methods that assess text-image alignment only at the final output, our approach provides

dynamic feedback during the diffusion pipeline, ensuring intermediate results remain semantically consistent with the input description. At step t, the latent representation z_t is decoded into the generation result $\tilde{x}_t = \mathcal{D}(z_t)$ via VAE decoder and optimize its CLIP-space distance to the target text-image pair:

$$\mathcal{L}_{tif} = 1 - \mathcal{S}\big(\Phi_{\mathcal{T}}(c'), \Phi_{\mathcal{I}}(\tilde{x}_t)\big), \tag{2}$$

here, $\Phi_{\mathcal{I}}(\cdot)$ denote the CLIP image encoder. $\mathcal{S}(\cdot)$ represents cosine similarity function.

Cross-Modal Multi-Control Adapter

RS T2I generation faces unique challenges: (1) Complex Spatial Control Integration: Users often provide diverse, potentially conflicting spatial controls that overlap in intricate geographical regions (e.g., mountainous cities), making it difficult to prioritize local signals. (2) Hierarchical Spatial Structures: RSIs exhibit strong multi-scale characteristics (e.g., cities, roads, buildings), requiring guidance at different semantic levels. To address these, we propose CMMCA that harmonizes any combination of spatial controls with enriched textual guidance for RS T2I generation. The proposed CMMCA using three token representations: (1) enriched text embedding $\mathcal{T}_e$, encoded from input text $\boldsymbol{T}$ via the CLIP text encoder; (2) query tokens $\boldsymbol{Q}$, implemented as trainable vectors that enable dynamic cross-modal interaction; and (3) vision tokens $\boldsymbol{V}$, extracted by a pre-trained visual encoder from user-provided image-based spatial controls.

Cross-Modal Context Aggregator. To resolve the ambiguity in prioritizing conflicting spatial control signals within overlapping geographical regions, we propose a Cross-Modal Context Aggregator. This module leverages the enriched text prompt as global semantic guidance to harmonize the multi-control inputs, ensuring that the generated scene adheres to the overall geographical context described in the text. In this module, a task-specific text embedding is first appended to the user-provided description before tokenization, effectively bridging the modality gap across diverse spatial conditions by embedding task-aware semantics into the textual representation. Subsequently, the query tokens $\boldsymbol{Q}$ and text embedding $\boldsymbol{T}$ are then concatenated and processed via self-attention at k-th Cross-Modal Context Aggregator:

$$[\boldsymbol{Q}_{k+1}, \boldsymbol{T}_{k+1}] = SelfAtt.([\boldsymbol{Q}_k, \boldsymbol{T}_k]). \tag{3}$$

The Self-attention enables query tokens to align semantically with user-control prompts by exchanging context with text embedding. This cross-modal aggregator enables dynamic alignment and semantic grounding between visual control signals and textual context, ensuring that the generated content remains globally coherent and contextually consistent under complex multi-control settings.

Multi-Control Cross Attention. To ensure that all input control conditions are semantically aligned and mutually compatible, we employ Multi-Control Cross Attention that facilitates interactions between the query tokens $\boldsymbol{Q}$ and the vision tokens $\boldsymbol{V}$ from all spatial control conditions.

Given n spatial conditions in diverse modalities, we first extract vision tokens $\boldsymbol{V}_{l,k}$ for each condition $\boldsymbol{C}_l$ from the k-th block of a pre-trained visual encoder. Let $\{\boldsymbol{V}_{1,k}, \boldsymbol{V}_{2,k}, ..., \boldsymbol{V}_{n,k}\}$ denote the collection of these vision tokens across all spatial conditions at k-th block. The process within the Multi-Control Cross Attention can be described as:

$$\mathcal{V}'_{l,k} = \mathcal{V}_{l,k} + \mathcal{P}, \tag{4}$$

$$\boldsymbol{Q}_j = CrossAtt.\left(\boldsymbol{Q}_j, [\mathcal{V}'_{1,j}, \mathcal{V}'_{2,j}, \ldots, \mathcal{V}'_{n,j}]\right), \tag{5}$$

where $\mathcal{P}$ denotes a learnable position embedding added to each $\boldsymbol{V}_{l,k}$, resulting in $\boldsymbol{V'}_{l,k}$, which facilitates better alignment between query and vision tokens. As this, the spatial conditions information embedded in the $\boldsymbol{V}$ is transferred to the $\boldsymbol{Q}$.

To ensure effective information interaction and compatibility among various control conditions, we introduce L alternating Cross-Modal Context Aggregator and Multi-Control Cross Attention modules. Furthermore, considering the inherent multi-scale nature of RS scenes (Zhang, Rao, and Agrawala 2023), we adopt a multi-level vision token extraction strategy. In each training step, visual tokens are extracted from different layers of the pre-trained visual encoder, corresponding to different semantic levels. These multi-level tokens from any provided spatial control conditions are fed into the Multi-Control Cross Attention module. This enables hierarchical modeling and guidance of spatial structures, ensuring both global layout coherence and local detail fidelity in the generated RSI.

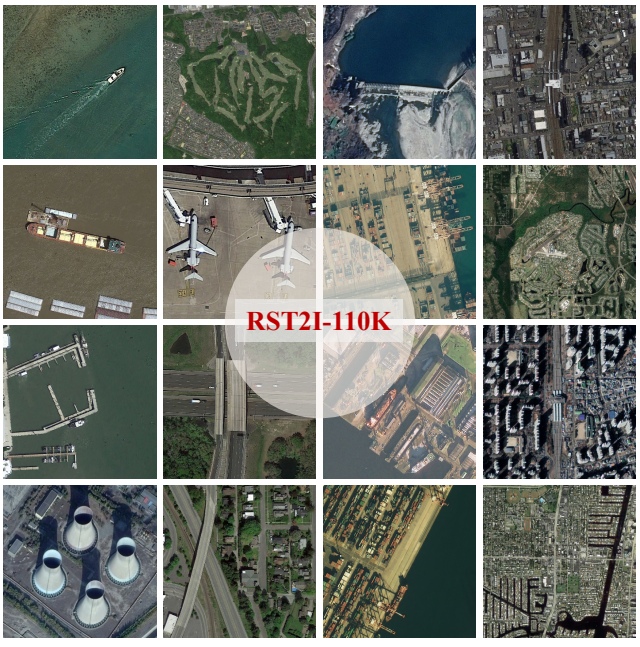

Figure 3: Examples of images from the RST2I-110K dataset.

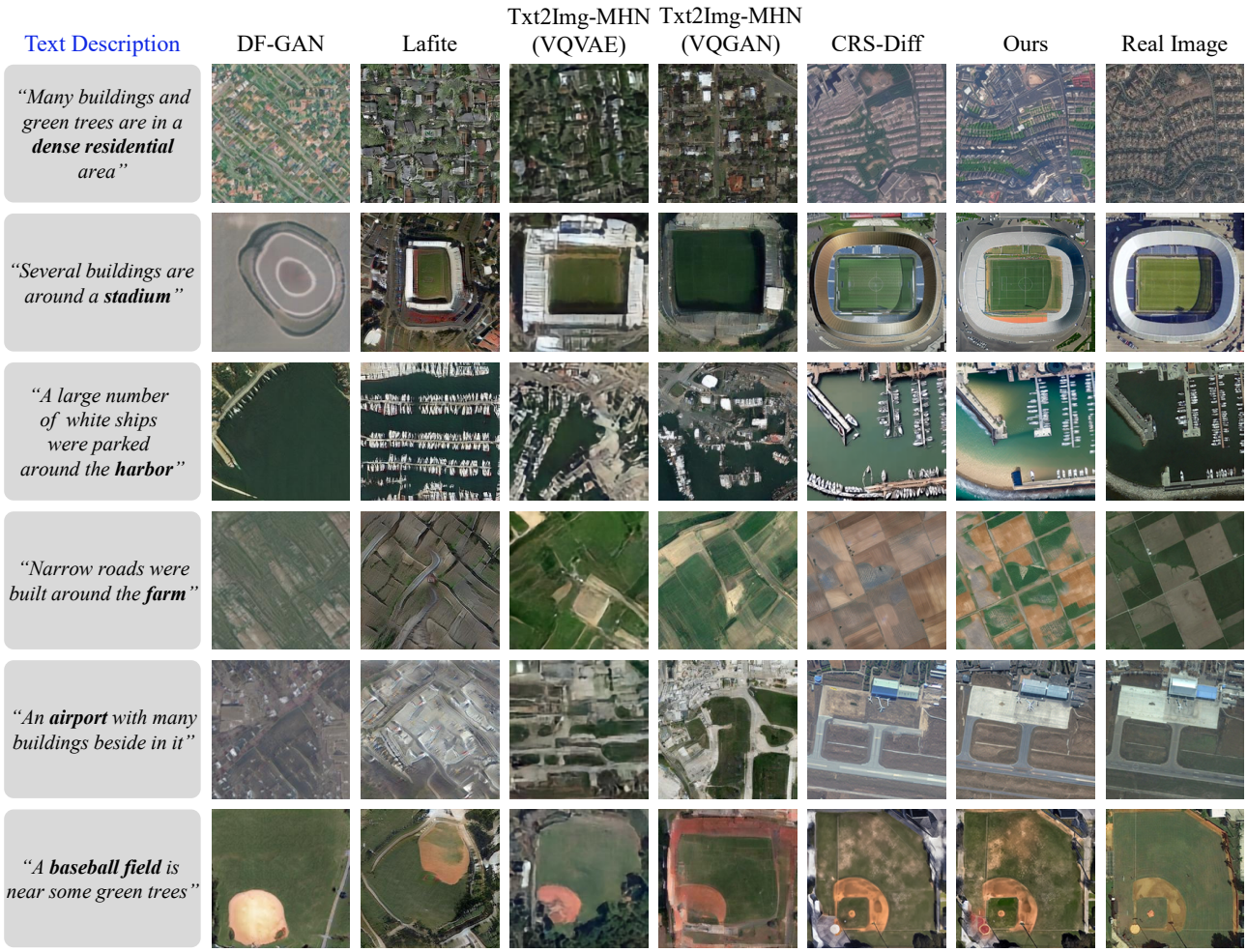

Figure 4: Visual comparison of our Any2RSI against other T2I methods on the RSICD test set.

Experiments

Experimental Setup

Datasets. We evaluate our method on both RSICD and RST2I-110K datasets. RSICD contains 10,921 images across 30 scene categories, each with five human-written captions, using 8,734 pairs for training and 2,187 for testing. To address limitations in scale and diversity, we build RST2I-110K by integrating and reprocessing three authoritative RS datasets: DOTA-V2.0 (Ding et al. 2021), DIOR (Li et al. 2020), and FAST (Wang et al. 2023) (derived from FAIR1M-2.0 (Sun et al. 2022)). The final dataset contains 115,200 curated image-text pairs (108,000 training, 7,200 test) with greater scene, object, and condition diversity. Each image is paired with one high-quality description generated by InternVL2.5-8B (Chen et al. 2024).

Evaluation Metrics. For text-to-image generation evaluation, we use four standard metrics: **1)** Fréchet Inception Distance (FID) (Heusel et al. 2017), **2)** Inception Score (IS) (Salimans et al. 2016), **3)** CLIP Score (Radford et al. 2021), and **4)** Zero-Shot Classification Overall Accuracy (OA) (Xu et al. 2023). Following Txt2Img-MHN, FID and IS are computed using an Inception-v3 model (Szegedy et al. 2016) fine-tuned on RSICD. The CLIP Score is evaluated using a CLIP model similarly fine-tuned on the same dataset.

Implementation Details. We adopt Stable Diffusion v1.5 as the backbone of Any2RSI. The model is fine-tuned for 40K iterations using the AdamW optimizer with a learning rate of 1×10^{-5}. For inference, we use the DDIM sampler with 50 denoising steps and a classifier-free guidance scale of 7.5. Following ControlNet, input images are converted into multiple conditional representations, such as segmentation masks, Canny edges, and HED maps (Xie and Tu 2015), using pre-trained annotator networks. The CMMCA leverages 256 query tokens with shared position embeddings across spatial conditions for fine-grained feature extraction, initialized from pre-trained QFormer weights (Li et al. 2023a), while the queries and embeddings are randomly initialized. All input and conditional images are resized to 512×512 pixels. Experiments are conducted on NVIDIA GeForce RTX 4090 GPUs with a batch size of 2.

Method	Source	Architecture	FID↓	Inception Score↑	CLIP Score↑	Zero-Shot OA↑
Attn-GAN (Xu et al. 2018)	CVPR'18	GAN-based	95.81	11.71	20.19	32.46
DAE-GAN (Ruan et al. 2021)	ICCV'21		93.15	7.71	19.69	29.74
DF-GAN (Tao et al. 2022)	CVPR'22		109.41	9.51	19.76	51.99
Lafite (Zhou et al. 2022)	CVPR'22		74.11	10.70	22.52	49.37
DALL-E (Ramesh et al. 2021)	ICML'21	Transformer-based	191.93	2.59	20.13	28.59
Txt2Img-MHN (VQVAE) (Xu et al. 2023)	TIP'23		175.36	3.51	21.35	41.46
Txt2Img-MHN (VQGAN) (Xu et al. 2023)	TIP'23		102.44	5.99	20.27	65.72
SD 1.5 (Rombach et al. 2022)	CVPR'22	Diffusion-based	113.25	12.31	20.04	51.96
SD 2.1 (Rombach et al. 2022)	CVPR'22		91.27	12.84	20.41	54.25
GLIGEN (Li et al. 2023b)	CVPR'23		80.11	13.52	19.87	57.76
ControlNet (Zhang, Rao, and Agrawala 2023)	ICCV'23		66.57	14.15	20.96	61.47
Uni-ControlNet (Zhao et al. 2023)	NeurIPS'23		61.12	15.53	20.74	64.01
ControlNet++ (Li et al. 2024a)	ECCV'24		51.23	16.18	20.92	68.33
CRS-Diff (Tang et al. 2024)	TGRS'24		53.49	16.07	21.05	67.82
EasyControl (Zhang et al. 2025c)	ICCV'25		48.71	16.37	21.49	71.28
Any2RSI (w/o EDG)	—		44.05	16.54	21.63	71.89
Any2RSI (Ours)	—		**41.14**	**16.86**	**21.87**	**72.54**

Table 1: Comparison of Any2RSI against baseline methods using four metrics: FID, Inception Score (IS), CLIP Score, and Zero-Shot Classification OA on the RSICD test set. Any2RSI (w/o EDG) indicates that EDG is not used, and the original coarse-grained texts are directly used as input. For fairness, all Diffusion-based methods have been retrained.

Method	HED	Canny	Seg.	HED+Canny	HED+Seg.	Canny+Seg.	HED+Canny+Seg.
	FID↓/SSIM↑	FID↓/SSIM↑	FID↓/mIoU↑	FID↓	FID↓/mIoU↑	FID↓/mIoU↑	FID↓/mIoU↑
SD 1.5	119.33/0.325	113.25/0.604	128.74/0.281	-	-	-	-
SD 2.1	100.39/0.357	91.27/0.629	107.98/0.309	-	-	-	-
GLIGEN	84.27/0.395	80.11/0.619	87.65/0.334	-	-	-	-
ControlNet	68.33/0.417	66.57/0.737	72.64/0.359	-	-	-	-
Uni-ControlNet	66.05/0.431	64.12/0.747	69.39/0.370	62.61	65.73/0.391	63.04/0.380	61.12/0.406
ControlNet++	62.58/0.466	56.34/0.762	67.02/0.357	56.83	57.45/0.417	61.36/0.395	57.38/0.414
CRS-Diff	59.23/0.479	58.15/0.751	65.18/0.354	53.49	59.91/0.426	56.28/0.410	54.16/0.403
EasyControl	51.36/0.473	48.57/0.814	50.13/**0.371**	50.79	49.61/**0.428**	50.85/**0.438**	47.91/0.438
Any2RSI (Ours)	**43.75/0.491**	**41.52/0.824**	**46.25**/0.365	**42.26**	**44.07**/0.410	**43.51**/0.427	**41.14/0.451**

Table 2: Results comparison of Any2RSI against other controllable T2I methods on the RSICD test set.

Method	FID↓	Inception Score↑	CLIP Score↑
GLIGEN	65.17	14.59	20.64
ControlNet	55.33	15.68	21.29
Uni-ControlNet	50.62	15.94	21.51
ControlNet++	49.28	16.13	21.13
CRS-Diff	46.61	16.42	21.34
EasyControl	41.03	16.90	21.81
Any2RSI (Ours)	**35.12**	**17.48**	**22.37**

Table 3: Results comparison of Any2RSI against baseline methods on the RST2I-110K test set with enriched text.

Controllable Text-to-Image Generation Results

Quantitative Analysis. As shown in Table 1, Any2RSI achieves the best performance on all evaluated metrics, particularly with an FID of 41.14 and a zero-shot overall accuracy (OA) of 72.54%, outperforming existing state-of-the-art methods. Diffusion-based models (e.g., ControlNet, Uni-ControlNet, CRS-Diff, and EasyControl) generally perform well, reflecting their suitability for high-resolution and semantically coherent generation. Transformer-based models like DALL-E and Txt2Img variants also achieve competitive results. In contrast, GAN-based methods (Attn-GAN, DAE-GAN, DF-GAN, and Lafite) underperform across most metrics, likely due to training instability and mode collapse. Moreover, a clear correlation exists between CLIP Score and Zero-Shot OA, suggesting CLIP Score serves as a reliable proxy for evaluating cross-modal consistency. Furthermore, under multi-condition settings (Table 2), Any2RSI achieves nearly the best performance across all metrics.

In addition to validation on RSICD, we evaluate Any2RSI on the proposed RST2I-110K dataset. As shown in Table 3, Any2RSI achieves the best FID score of 35.12, outperforming ControlNet++ at 49.28, CRS-Diff at 48.61, and EasyControl at 41.03, which indicates better preservation of global structures and fine-grained details in complex scenes. It also attains the highest Inception Score of 17.48, reflect-

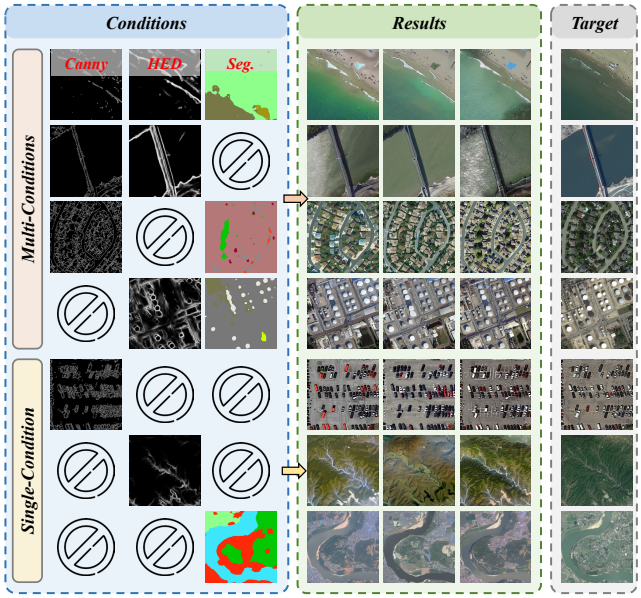

Figure 5: Visual comparison under multi-conditions and single-condition settings.

ing strong visual fidelity and generative diversity. Notably, Any2RSI achieves the best CLIP Score of 22.37, demonstrating superior text-image alignment that is critical for accurate semantic grounding in RS applications such as urban planning and disaster simulation.

Qualitative Analysis. As shown in Figure 4, Any2RSI achieves superior performance in RS image generation, particularly in semantic fidelity, structural coherence, and visual realism. Qualitative comparisons show that Any2RSI-generated images closely resemble the ground truth across diverse complex scenes, including dense residential areas, stadiums, harbors, farm, airports, and baseball fields. The effectiveness of Any2RSI stems from its integration of enriched textual descriptions and multiple control conditions. By leveraging detailed captions with object-level information and spatial relationships, the model receives stronger semantic guidance than standard text inputs, enabling better understanding of fine-grained scene semantics and more accurate outputs. Furthermore, as shown in Figure 5, we present the generation results under different single-condition and multi-condition controls, with three images generated each time. The results demonstrate that our method can maintain excellent and flexible generation performance under various condition combinations.

Ablation Study

Effects of Different Components. As shown in Table 4, ablation studies on Any2RSI start from a baseline with an FID of 58.95 and an Inception Score of 15.02. Adding any single component improves performance, and CMMCA yields the strongest gain with an FID of 51.50 and an Inception Score of 16.30. Combining components further enhances results, and the full model achieves the best performance with an FID of 41.14 and an Inception Score of 16.86, demonstrating clear synergy among EDG, PGO, and CMMCA.

Index	EDG	PGO	CMMCA	FID↓	IS↑
Baseline	✗	✗	✗	58.95	15.02
I	✓	✗	✗	53.13	15.87
II	✗	✓	✗	54.67	16.09
III	✗	✗	✓	51.50	16.30
IV	✓	✓	✗	47.60	16.23
V	✓	✗	✓	45.52	16.67
VI	✗	✓	✓	44.05	16.54
Ours	✓	✓	✓	**41.14**	**16.86**

Table 4: Effectiveness of each component.

Method	HED		Canny		Seg.	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑
w/o EDG	49.03	16.18	45.52	16.67	53.28	15.97
GPT-4o	45.23	16.71	42.69	15.48	49.17	16.25
LLaVA-1.5-7B	46.26	**16.97**	43.09	16.33	50.17	15.92
Qwen2.5-VL-7B	**44.15**	16.87	42.03	**17.24**	47.38	**16.31**
InternVL2.5-8B	44.35	16.62	**41.52**	16.81	**46.25**	16.13

Table 5: Effectiveness of various VLMs.

Effects on Different Vision-Language Models. We systematically evaluate the impact of different Vision-Language Models (VLMs) on RS T2I generation by using their outputs as textual conditions. As shown in Table 5, the choice of VLM influences generation quality. Under HED guidance, Qwen2.5-VL-7B (Bai et al. 2025) achieves the best FID of 44.15, while LLaVA-1.5-7B (Liu et al. 2024) obtains the highest Inception Score of 16.97. In the Canny condition, InternVL2.5-8B (Chen et al. 2024) yields the best FID of 41.52, and Qwen2.5-VL-7B attains the highest Inception Score of 17.24. For segmentation, all models exhibit degraded performance, which is consistent with the limited spatial guidance provided by segmentation maps in relatively simple scenes. Notably, all VLM-based methods outperform the baseline (w/o EDG), confirming their effectiveness in enhancing text quality for RS T2I generation.

Conclusion

In this paper, we propose Any2RSI, a controllable RS T2I generation method that improves con- trol flexibility and semantic consistency. First, we design the CMMCA to unify heterogeneous spatial conditions. Second, we introduce EDG module to enrich textual prompts via cross-modal semantics. Finally, we present RST2I-110K, a large-scale dataset of over 115,000 high-quality image-text pairs across diverse scenes, expanding resources for RS T2I generation research. Extensive experiments show that Any2RSI achieves state-of-the-art performance on both existing and new datasets, improving the realism and structural accuracy of generated RS imagery.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62431020, the National Key Research and Development Program of China under Grant 2024YFE0111800, and the Fundamental Research Funds for the Central Universities under Grant 2042025kf0030.

References

Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; and Yin, X. 2023. Spatext: Spatio-textual representation for controllable image generation. In *CVPR*, 18370–18380.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, X.; Zhang, Z.; Zhang, H.; Zhou, Y.; Kim, S. Y.; Liu, Q.; Li, Y.; Zhang, J.; Zhao, N.; Wang, Y.; et al. 2025. Unireal: Universal image generation and editing via learning real-world dynamics. In *CVPR*, 12501–12511.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; and Wang, Z. 2022. NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning. *IEEE TGRS*, 60: 1–19.

Ding, J.; Xue, N.; Xia, G.-S.; Bai, X.; Yang, W.; Yang, M. Y.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; et al. 2021. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE TPAMI*, 44(11): 7778–7796.

Goktepe, M.; hossein Shamseddin, A.; Uysal, E.; Monteagudo, J. M.; Drees, L.; Toker, A.; Asseng, S.; and von Bloh, M. 2025. EcoMapper: Generative Modeling for Climate-Aware Satellite Imagery. In *ICML*, 19734–19754.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 6629–6640.

Himeur, Y.; Rimal, B.; Tiwary, A.; and Amira, A. 2022. Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives. *IF*, 86: 44–75.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*, 6840–6851.

Huang, V. S.-J.; Zhuo, L.; Xin, Y.; Wang, Z.; Wang, F.-Y.; Wang, Y.; Zhang, R.; Gao, P.; and Li, H. 2025. Tide: Temporal-aware sparse autoencoders for interpretable diffusion transformers in image generation. *arXiv preprint arXiv:2503.07050*.

Khanna, S.; Liu, P.; Zhou, L.; Meng, C.; Rombach, R.; Burke, M.; Lobell, D.; and Ermon, S. 2024. DiffusionSat: A Generative Foundation Model for Satellite Imagery. In *ICLR*, 5586–5604.

Li, J.; He, W.; Li, Z.; Guo, Y.; and Zhang, H. 2025a. Overcoming the uncertainty challenges in detecting building changes from remote sensing images. *ISPRS P&RS*, 220: 1–17.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742.

Li, K.; Cao, X.; Deng, Y.; Pang, C.; Xin, Z.; Meng, D.; and Wang, Z. 2025b. Dynamicearth: How far are we from open-vocabulary change detection? *arXiv preprint arXiv:2501.12931*.

Li, K.; Wan, G.; Cheng, G.; Meng, L.; and Han, J. 2020. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS P&RS*, 159: 296–307.

Li, M.; Yang, T.; Kuang, H.; Wu, J.; Wang, Z.; Xiao, X.; and Chen, C. 2024a. ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback. In *ECCV*, 129–147.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023b. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 22511–22521.

Li, Z.; Cheng, T.; Chen, S.; Sun, P.; Shen, H.; Ran, L.; Chen, X.; Liu, W.; and Wang, X. 2025c. ControlAR: Controllable Image Generation with Autoregressive Models. In *ICLR*, 8450–8474.

Li, Z.; He, W.; Li, J.; Lu, F.; and Zhang, H. 2024b. Learning without exact guidance: Updating large-scale high-resolution land cover maps from low-resolution historical labels. In *CVPR*, 27717–27727.

Liu, C.; Chen, K.; Zhao, R.; Zou, Z.; and Shi, Z. 2025. Text2Earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. *IEEE GRSM*, 13(3): 238–259.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *CVPR*, 26296–26306.

Lu, W.; Chen, S.-B.; Li, H.-D.; Shu, Q.-L.; Ding, C. H. Q.; Tang, J.; and Luo, B. 2025. LEGNet: A Lightweight Edge-Gaussian Network for Low-Quality Remote Sensing Image Object Detection. In *ICCVW*, 2844–2853.

Lu, W.; Chen, S.-B.; Shu, Q.-L.; Tang, J.; and Luo, B. 2024. DecoupleNet: A Lightweight Backbone Network With Efficient Feature Decoupling for Remote Sensing Visual Tasks. *IEEE TGRS*, 62: 1–13.

Lu, W.; Chen, S.-B.; Tang, J.; Ding, C. H.; and Luo, B. 2023. A robust feature downsampling module for remote-sensing visual tasks. *IEEE TGRS*, 61: 1–12.

Lu, X.; Wang, B.; Zheng, X.; and Li, X. 2017. Exploring models and data for remote sensing image caption generation. *IEEE TGRS*, 56(4): 2183–2195.

Pan, Y.; He, Q.; Jiang, Z.; Xu, P.; Wang, C.; Peng, J.; Wang, H.; Cao, Y.; Gan, Z.; Chi, M.; et al. 2025. Pixelponder: Dynamic patch adaptation for enhanced multi-conditional text-to-image generation. *arXiv preprint arXiv:2503.06684*.

Pang, L.; Cao, X.; Tang, D.; Xu, S.; Bai, X.; Zhou, F.; and Meng, D. 2025. HSIGene: a Foundation Model for Hyperspectral Image Generation. *IEEE TPAMI*, 1–18.

Qin, C.; Zhang, S.; Yu, N.; Feng, Y.; Yang, X.; Zhou, Y.; Wang, H.; Niebles, J. C.; Xiong, C.; Savarese, S.; Ermon, S.; Fu, Y.; and Xu, R. 2023. UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild. In *NeurIPS*, 42961–42992.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *ICML*, 8821–8831.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *ICML*, 1060–1069.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Rong, F.; Lan, M.; Zhang, Q.; and Zhang, L. 2025. Customized SAM 2 for Referring Remote Sensing Image Segmentation. *arXiv preprint arXiv:2503.07266*.

Ruan, S.; Zhang, Y.; Zhang, K.; Fan, Y.; Tang, F.; Liu, Q.; and Chen, E. 2021. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *ICCV*, 13960–13969.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *NeurIPS*, 2234–2242.

Sastry, S.; Khanal, S.; Dhakal, A.; and Jacobs, N. 2024. Geosynth: Contextually-aware high-resolution satellite image synthesis. In *CVPRW*, 460–470.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising diffusion implicit models. In *ICLR*.

Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. 2022. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS P&RS*, 184: 116–130.

Sun, Y.; Liu, Y.; Tang, Y.; Pei, W.; and Chen, K. 2024. Anycontrol: create your artwork with versatile control on text-to-image generation. In *ECCV*, 92–109.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2025. OminiControl: Minimal and Universal Control for Diffusion Transformer. In *ICCV*, 14940–14950.

Tang, D.; Cao, X.; Hou, X.; Jiang, Z.; Liu, J.; and Meng, D. 2024. CRS-Diff: Controllable Remote Sensing Image Generation With Diffusion Model. *IEEE TGRS*, 62: 1–14.

Tao, M.; Tang, H.; Wu, F.; Jing, X.-Y.; Bao, B.-K.; and Xu, C. 2022. Df-gan: A simple and effective baseline for text-to-image synthesis. In *CVPR*, 16515–16525.

Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; and Zhang, L. 2023. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. In *NeurIPS*, 8815–8827.

Wang, H.; Peng, J.; He, Q.; Yang, H.; Jin, Y.; Wu, J.; Hu, X.; Pan, Y.; Gan, Z.; Chi, M.; Peng, B.; and Wang, Y. 2025. UniCombine: Unified Multi-Conditional Combination with Diffusion Transformer. In *ICCV*, 18325–18334.

Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Li, C.; Wang, S.; Huang, T.; and Liu, Z. 2025. Omnigen: Unified image generation. In *CVPR*, 13294–13304.

Xie, S.; and Tu, Z. 2015. Holistically-nested edge detection. In *ICCV*, 1395–1403.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, 1316–1324.

Xu, Y.; Wang, D.; Zhang, L.; and Zhang, L. 2025. Dual Selective Fusion Transformer Network for Hyperspectral Image Classification. *NN*, 187: 107311.

Xu, Y.; Yu, W.; Ghamisi, P.; Kopp, M.; and Hochreiter, S. 2023. Txt2Img-MHN: Remote sensing image generation from text using modern Hopfield networks. *IEEE TIP*, 32: 5737–5750.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 5907–5915.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*, 3836–3847.

Zhang, X.; Cai, N.; Zhang, H.; Zhang, Y.; Di, J.; and Lin, W. 2023. AFD-Former: A Hybrid Transformer With Asymmetric Flow Division for Synthesized View Quality Enhancement. *IEEE TCSVT*, 33(8): 3786–3798.

Zhang, X.; Ma, J.; Wang, G.; Zhang, Q.; Zhang, H.; and Zhang, L. 2025a. Perceive-IR: Learning to Perceive Degradation Better for All-in-One Image Restoration. *IEEE TIP*.

Zhang, X.; Zhang, H.; Wang, G.; Zhang, Q.; Zhang, L.; and Du, B. 2025b. UniUIR: Considering Underwater Image Restoration as an All-in-One Learner. *IEEE TIP*, 34: 6963–6977.

Zhang, Y.; Yuan, Y.; Song, Y.; Wang, H.; and Liu, J. 2025c. EasyControl: Adding Efficient and Flexible Control for Diffusion Transformer. In *ICCV*, 19513–19524.

Zhang, Z.; Zhao, T.; Guo, Y.; and Yin, J. 2024. RS5M and GeoRSCLIP: A Large-Scale Vision- Language Dataset and a Large Vision-Language Model for Remote Sensing. *IEEE TGRS*, 62: 1–23.

Zhao, S.; Chen, D.; Chen, Y.-C.; Bao, J.; Hao, S.; Yuan, L.; and Wong, K.-Y. K. 2023. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 11127–11150.

Zhou, Y.; Zhang, R.; Chen, C.; Li, C.; Tensmeyer, C.; Yu, T.; Gu, J.; Xu, J.; and Sun, T. 2022. Towards language-free training for text-to-image generation. In *CVPR*, 17907–17917.