GeoCore-9B: Towards Geo-Aware Generative Foundation Models in Earth Observation

Jeonghyeok Do
Information & Electronics Research Institute
KAIST
ehwjdgur0913@kaist.ac.kr

Munchurl Kim*
School of Electrical Engineering
KAIST
mkimee@kaist.ac.kr

Project page: https://kaist-viclab.github.io/GeoCore-9B_site/

Abstract

Existing generative models for earth observation (EO) predominantly rely on fine-tuning natural image priors, which limits their scalability and introduces perspective biases that conflict with geospatial constraints. To address this, we introduce **GeoCore-9B**, a 9-billion-parameter generative foundation model, which is the first of its scale to be trained from scratch exclusively on EO data. Unlike previous EO foundation models, GeoCore-9B is built upon a Flow Matching-based Diffusion Transformer (DiT) and natively conditions generation on text descriptions and continuous geospatial metadata, including ground sample distances, latitudes, and longitudes. To overcome the convergence and spatial disorientation challenges of training at this scale, we propose a Geospatial Semantic Alignment loss. This objective distills structural Earth surface priors (e.g., terrain and urban areas) from a frozen specialist teacher network, constraining the diffusion latent trajectory during training without adding inference overhead. Pre-trained on the global-scale Git-10M dataset, GeoCore-9B demonstrates strong downstream versatility. Beyond standard proxy generative tasks, we show that GeoCore-9B can be effectively adapted for practical EO applications, including highly challenging tasks such as cloud removal and SAR-to-optical cross-modal translation. Extensive evaluations confirm that GeoCore-9B establishes new state-of-the-art performance in both visual fidelity and geographic structural accuracy.

1 Introduction

Generative foundation models for Earth Observation (EO) [13, 16, 23, 27, 38, 41, 51] have emerged as a pivotal technology for simulating terrestrial environments, augmenting scarce datasets, and enabling complex downstream applications. Beyond generic image synthesis, a practically useful EO foundation prior should be transferable to restoration and cross-modal translation scenarios, where the models must recover geographically faithful structures from real occlusions, degradations, or sensor-induced modality gaps. However, generating satellite imagery faces unique challenges: unlike natural images, EO data is strictly orthographic, physically anchored by spatial resolution (ground sample distance, GSD), and covers highly dense and heterogeneous geographic structures across the globe. Therefore, establishing a robust generative world model that inherently comprehends these physical and geometric constraints remains an ongoing challenge.

Existing generative approaches [16, 23, 38, 41] in the EO domain have primarily relied on *fine-tuning U-Net-based pre-trained models* (e.g., Stable Diffusion v1.5 and v2.1 [35]) *originally optimized for natural images*. While this fine-tuning paradigm accelerates convergence, it inevitably introduces severe domain shifts. Natural image priors are inherently biased toward perspective projection,

*Corresponding author.

Preprint.

center-object framing, and casual spatial scales, which fundamentally conflict with the scale-invariant, bird's-eye view nature of satellite imagery. Furthermore, the reliance on standard U-Net architectures severely bottlenecks their representational capacity. Consequently, these models often struggle with geometric distortions and fail to capture authentic geospatial data distributions. Another limitation lies in how downstream capability has been evaluated. Existing evaluations often focus on controllability-oriented or proxy generative settings, such as sketch-conditioned EO generation [41] and multimodal generation [23] built by synthetically augmenting the RSICD [28] text-to-image benchmark. While useful for assessing conditional generation, such protocols provide limited evidence that the learned generative prior can transfer to practical EO tasks involving real paired observations, occlusions, degradations, or severe cross-sensor modality gaps. Even a recent attempt [13] to build an EO model from scratch remains constrained: its representational capacity is often diluted across broad multimodal tasks, and it is a relatively small-scale model, lacking the massive parameter scale required to synthesize the immense complexity of the Earth's surface.

To address these limitations, we propose **GeoCore-9B**, a 9-billion-parameter generative foundation model trained from scratch. GeoCore-9B is the first generative foundation model built upon a Flow Matching-based Diffusion Transformer (DiT) [6, 18, 30] for EO. It is pre-trained on the global-scale Git-10M [23] dataset and conditions generation on text descriptions and geospatial metadata, including GSD, latitude, and longitude. This design avoids reliance on natural image priors and enables geo-aware EO synthesis. Training such a large model from scratch, however, poses severe convergence and spatial disorientation challenges. To address this, we introduce a **Geospatial Semantic Alignment** loss, a training-only alignment objective that distills satellite-specific structural cues from a frozen DINOv3-Sat [39] as a teacher network. By aligning intermediate DiT representations with geospatial semantic features, GeoCore-9B improves structural fidelity in generated EO images without adding inference overhead. Beyond pre-training, we validate the transferability of the learned EO foundation prior on practical downstream applications. For example, even with parameter-efficient fine-tuning (e.g., LoRA [9]), GeoCore-9B can be adapted to highly challenging tasks such as cloud removal and SAR-to-optical translation, demonstrating its efficacy and utility beyond controllability-oriented or synthetic proxy generation tasks. The main contributions of our work are summarized as follows:

- We introduce **GeoCore-9B**, a 9-billion-parameter DiT-based generative foundation model trained from scratch on EO data with text and geospatial metadata.
- We propose a **Geospatial Semantic Alignment** loss, which uses a frozen satellite-specialist teacher to improve structural fidelity during training with zero inference overhead.
- We demonstrate practical downstream transferability by adapting GeoCore-9B to cloud removal and SAR-to-optical translation, where it outperforms or remains competitive compared to task-specific specialist methods.

2 Related Work

Generative models in Earth Observation (EO). Generative modeling for EO [13, 16, 23, 27, 38, 41, 51] has evolved from adopting natural-image diffusion models to developing domain-specific and multimodal frameworks. DiffusionSat [16] adapts U-Net-based pre-trained models [35] by incorporating temporal and multi-spectral conditions for satellite image generation. RS-Diff [38] proposes a cascaded architecture that sequentially generates and super-resolves remote sensing imagery. CRS-Diff [41] improves controllability by injecting composite spatial signals, such as sketches and semantic masks, through multi-scale feature fusion. Text2Earth [23] scales text-driven EO generation with the global-scale Git-10M dataset, while TerraMind [13] introduces an any-to-any multimodal framework with a unified transformer backbone. Our proposed GeoCore-9B further scales generative pre-training from scratch on EO data and incorporates geospatial metadata for geo-aware synthesis.

Scalable generative architectures and alignment. Large-scale image generation has shifted from U-Net-based diffusion models [35] to Diffusion Transformers (DiT) [30], as demonstrated by recent Flow Matching-based models such as Stable Diffusion v3.0 [6] and FLUX [18]. Flow Matching [22, 25] learns a continuous vector field from noise to data, offering a scalable and stable training objective for large generative models. In parallel, representation alignment has been shown to improve diffusion training by aligning generative features with strong visual representations [20, 47, 50]. We build

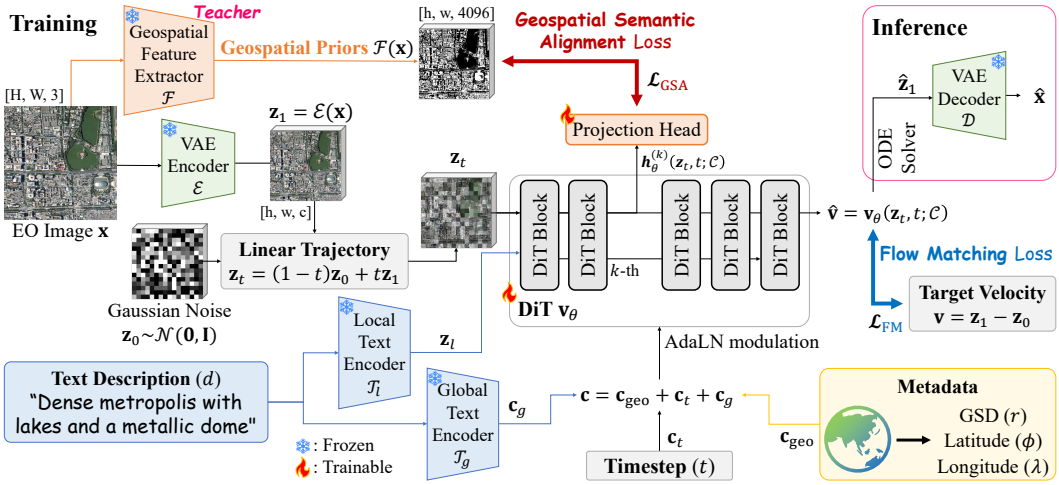

Figure 1: Overview of GeoCore-9B. GeoCore-9B trains a Flow Matching-based DiT in the latent space, conditioned on local text tokens and global geospatial metadata through token fusion and AdaLN. A training-only Geospatial Semantic Alignment loss distills structural Earth priors from a frozen satellite-specialist teacher, improving spatial fidelity without adding inference overhead.

our GeoCore-9B on these advances by combining a Flow Matching-based DiT backbone with a training-only geospatial alignment objective specifically tailored to satellite imagery.

3 GeoCore-9B

Fig. 1 illustrates the conceptual flow of our GeoCore-9B built upon a Flow Matching-based DiT [6, 18, 30], augmented with text and geospatial metadata conditioning, alongside a training-only semantic alignment objective.

3.1 Geo-Conditioned Flow Matching Backbone

GeoCore-9B operates in the latent space of a pre-trained VAE [18]. Given an RGB image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, we obtain its latent representation $\mathbf{z}_1 = \mathcal{E}(\mathbf{x})$. Following Flow Matching [22, 25], we sample $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and define a linear trajectory:

$$\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1, \quad t \in [0,1], \tag{1}$$

where the target velocity is $\mathbf{v} = \mathbf{z}_1 - \mathbf{z}_0$. The DiT backbone $\mathbf{v}_\theta$ is trained to predict $\mathbf{v}$ from the noisy latent $\mathbf{z}_t$ under text and geospatial conditions. To condition generation on a text description d, we use a global text embedding $\mathcal{T}_g(d)$ from CLIP-ViT-L/14 [31] and a token-level text embedding $\mathcal{T}_l(d)$ from T5-XXL [32] as:

$$\mathbf{c}_g = \mathcal{T}_g(d) \in \mathbb{R}^{1\times c_g}, \qquad \mathbf{z}_l = \mathcal{T}_l(d) \in \mathbb{R}^{M\times c_l}, \tag{2}$$

where M is the text sequence length. After linear projection, $\mathbf{z}_l$ is concatenated with the image latent tokens $\mathbf{z}_t$ and processed by the DiT blocks, while $\mathbf{c}_g$ is used for global modulation. We use 3D rotary positional embeddings (RoPE) [40] to encode both text-token positions and latent spatial coordinates.

3.2 Geospatial Metadata Conditioning

Satellite RGB imagery is strongly tied to physical scale and geographic location. Given a GSD r, latitude ϕ, and longitude λ, we map each scalar with a sinusoidal projection $\Phi(\cdot)$ and combine them to obtain the geospatial context vector $\mathbf{c}_{\text{geo}}$ as:

$$\mathbf{c}_{\text{geo}} = \text{MLP}_r\left(\Phi(r)\right) + \text{MLP}_\phi\left(\Phi(\phi)\right) + \text{MLP}_\lambda\left(\Phi(\lambda)\right). \tag{3}$$

Then, we obtain a global conditioning vector $\mathbf{c}$ by adding $\mathbf{c}_{\text{geo}}$, the timestep embedding $\mathbf{c}_t$, and the global text embedding $\mathbf{c}_g$ as $\mathbf{c} = \mathbf{c}_{\text{geo}} + \mathbf{c}_t + \mathbf{c}_g$, where $\mathbf{c}$ modulates the DiT activations through AdaLN [30]. For classifier-free guidance [8], we apply condition dropout to the text and geospatial metadata, using learnable null embeddings for missing metadata conditions.

3.3 Geospatial Semantic Alignment

Training a 9B-parameter DiT from scratch on EO data is challenging due to slow convergence and spatially unstable generation. To stabilize training, we introduce a Geospatial Semantic Alignment (GSA) loss, a training-only feature alignment objective. We use a frozen DINOv3-Sat [39] encoder $\mathcal{F}(\cdot)$ as a satellite-specialist teacher and align intermediate DiT features with its dense structural representations. Given intermediate latent tokens $\mathbf{h}_\theta^{(k)}(\mathbf{z}_t, t, \mathcal{C})$ at layer k, the GSA loss is defined as:

$$\mathcal{L}_{\mathrm{GSA}} = \mathbb{E}_{\mathbf{z}_0,\mathbf{x},t,\mathcal{C}}\left[\left\|\mathbf{W}_{\mathrm{proj}}\left(\mathbf{h}_\theta^{(k)}(\mathbf{z}_t, t, \mathcal{C})\right) - \mathcal{F}(\mathbf{x})\right\|_2^2\right], \tag{4}$$

where $\mathcal{C} = \{d, r, \phi, \lambda\}$ is the conditioning set and $\mathbf{W}_{\mathrm{proj}}$ maps the DiT features to the teacher feature dimension. Since $\mathcal{F}$ and $\mathbf{W}_{\mathrm{proj}}$ are used only during training, the GSA loss improves structural fidelity without increasing inference cost.

3.4 Training Objective

The final training objective combines the Flow Matching loss ($\mathcal{L}_{\mathrm{FM}}$) and the GSA loss as:

$$\mathcal{L}_{\mathrm{total}} = \underbrace{\mathbb{E}_{\mathbf{z}_0,\mathbf{z}_1,t,\mathcal{C}}\left[\|\mathbf{v}_\theta(\mathbf{z}_t, t, \mathcal{C}) - (\mathbf{z}_1 - \mathbf{z}_0)\|_2^2\right]}_{\mathcal{L}_{\mathrm{FM}}} + \mu\mathcal{L}_{\mathrm{GSA}}, \tag{5}$$

where μ controls the power of the semantic alignment and is empirically set to 0.5 in our experiments.

4 Experiments

4.1 Datasets

We evaluate GeoCore-9B on both generative and practical downstream EO tasks. For pre-training, we use Git-10M [23], a global-scale satellite RGB image dataset containing 10M image-text pairs with geospatial metadata, including GSDs, latitudes, and longitudes. For text-to-image adaptation, we use RSICD [28], which contains 10,921 remote sensing image-text pairs across 30 scene categories but does not provide geospatial metadata. For practical downstream evaluation, we use Sen2-MTC [10] for cloud removal and QXS-SAROPT [11] for SAR-to-optical image translation.

4.2 Implementation Details

Pre-training. Input images are randomly cropped to $256 \times 256 \times 3$ and encoded into a $32 \times 32 \times 32$ latent space using a pre-trained VAE encoder [18]. After 2×2 patchification, the $16 \times 16 \times 128$ latent token grid is projected to a hidden dimension of $4,096$. GeoCore-9B uses 32 DiT blocks with 32 attention heads, an MLP ratio of 3.0, and 3D RoPE [40] axis dimensions of 32 (text), 48 (image height), and 48 (image width). Token-level text embeddings are truncated or padded to a maximum sequence length of $M = 256$ and have a dimension of $c_l = 4,096$, while the global text embedding has a dimension of $c_g = 768$; both are projected to the DiT hidden dimension before conditioning. The timestep, GSD, longitude, and latitude conditions are each encoded as 256-dimensional sinusoidal features, and are projected to the hidden dimension through separate MLP embedders. For the GSA loss, the frozen DINOv3-Sat [39] teacher produces $16 \times 16 \times 4,096$ dense features, and we apply the alignment loss at the $k = 8$-th DiT block after projecting the intermediate DiT features to the teacher feature space, with $\mu = 0.5$. We train GeoCore-9B from scratch on the full Git-10M dataset for 300K iterations using AdamW with a constant learning rate of 1×10^{-4}, a weight decay of 0.001, a global batch size of $1,024$, and bfloat16 (bf16) mixed precision. Following Text2Earth [23], we adopt a progressive data refinement strategy: after pre-training on the full Git-10M corpus, we further refine GeoCore-9B on the high-quality subset whose quality scores exceed 4.8 [23], improving visual fidelity and fine-grained details. Training uses DeepSpeed ZeRO-2 [33] and takes approximately 15 days on eight NVIDIA Blackwell B200 GPUs.

Downstream adaptation. For downstream tasks, we perform parameter-efficient adaptation by freezing the pre-trained GeoCore-9B backbone and optimizing lightweight LoRA adapters [9]. We use a rank of $r_{\mathrm{LoRA}} = 64$ and a scaling factor of $\alpha_{\mathrm{LoRA}} = 128$, injecting LoRA into the linear

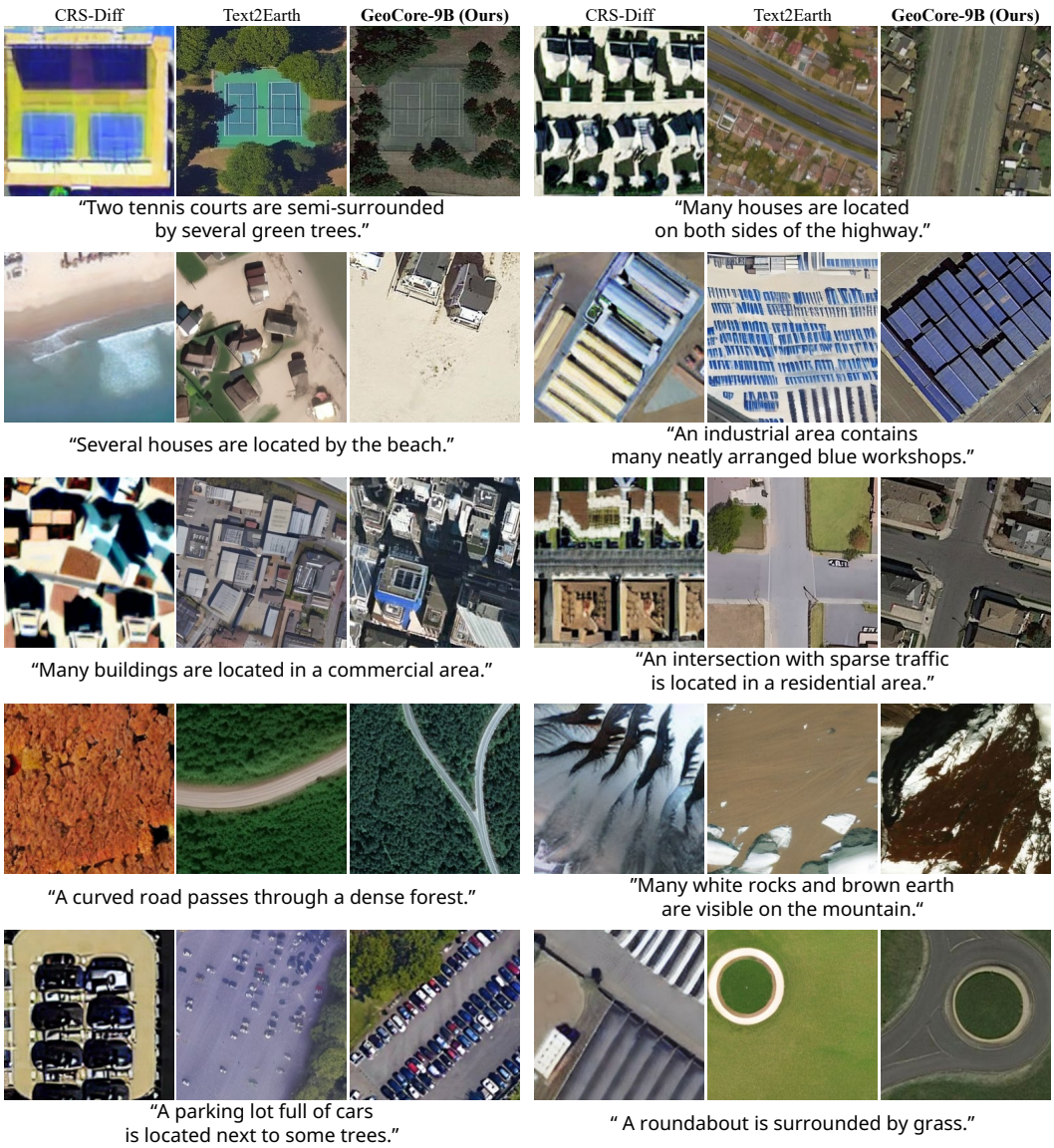

Figure 2: Qualitative comparison of text-conditioned generation. Given various text descriptions, GeoCore-9B synthesizes highly realistic and structurally accurate satellite imagery, outperforming baselines (CRS-Diff [41] and Text2Earth [23]) which often suffer from severe artifacts.

layers of the attention and feed-forward modules. The adapters are optimized with a learning rate of 2×10^{-4} and a batch size of 256. For image-conditioned tasks, the condition images and target RGB images are encoded by the frozen VAE encoder, and the condition latent is concatenated with the noisy target latent $\mathbf{z}_t$ before the first input projection layer. We additionally optimize this input projection layer to accommodate the enlarged conditional latent input.

Inference. We use a first-order Euler solver for the Flow Matching ODE [22, 25] with 50 sampling steps. Classifier-free guidance is applied with a scale of $w = 4.0$, using an empty text prompt for text and learned null embeddings for geospatial metadata.

4.3 Zero-shot Image Generation

We first evaluate GeoCore-9B without task-specific fine-tuning to examine whether pre-training learns geo-aware generative priors. Given a text description and geospatial metadata, the model generates RGB images conditioned on semantic contents, physical scales, and geographic locations.

Figure 3: Qualitative comparison across varying ground sample distances (GSD). GeoCore-9B adaptively adjusts visual granularity from fine structural details ($1m$) to broad land-cover patterns ($32m$), demonstrating superior scale-awareness compared to CRS-Diff [41] and Text2Earth [23] which struggle with unnatural textures and scale inconsistency.

Text. To evaluate semantic controllability, we vary the text prompts while fixing the geospatial metadata. As shown in Fig. 2, GeoCore-9B accurately follows prompts describing diverse scenes (e.g., residential areas, coastal regions, and industrial zones) while maintaining the authentic orthographic structures of satellite imagery. In contrast, baseline models such as CRS-Diff [41] and Text2Earth [23] often suffer from severe structural artifacts or unnatural textures.

GSD. To assess scale controllability, we vary the GSD values while fixing the text prompts and geographic coordinates. As illustrated in Fig. 3, GeoCore-9B adaptively adjusts visual granularity according to the physical resolution. It produces finer structural details at lower GSD values (e.g., 1m) and coarser, broader land-cover patterns at higher GSD values (e.g., 32m), demonstrating superior scale-awareness compared to the baselines [23, 41].

Geographic coordinates. To rigorously evaluate geographic conditioning, we challenge the models with a strictly text-free input setting: latitude and longitude coordinates *only*. As shown in Fig. 4, while the baseline model (CRS-Diff [41]) fails to generate meaningful imagery without text prompts, suffering from severe artifacts and repeating patterns, GeoCore-9B successfully retrieves location-dependent geographic priors. It synthesizes highly accurate terrains corresponding precisely to the given coordinates.

4.4 Ablation Study

We evaluate the effect of Geospatial Semantic Alignment (GSA) by training a variant without the GSA objective, i.e., $\mu = 0$. As shown in Fig. 5, removing GSA leads to fragmented textures and distorted

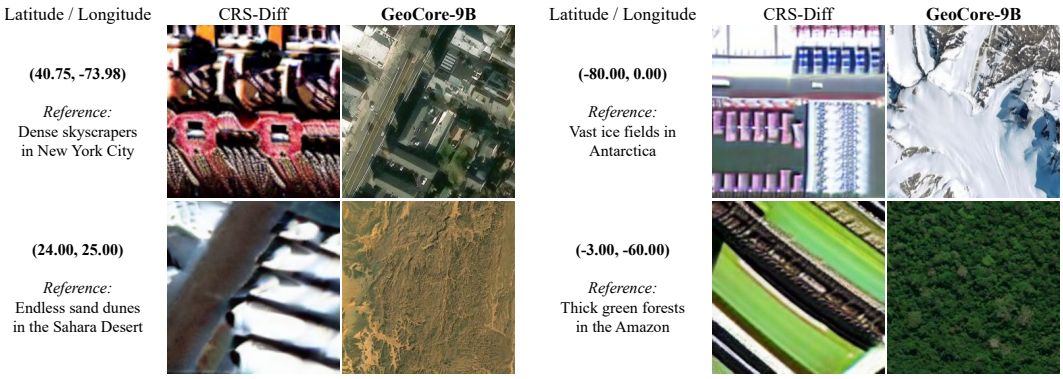

Figure 4: Qualitative comparison of text-free generation guided solely by latitude and longitude coordinates. Without text prompts, the baseline model (CRS-Diff [41]) fails to generate meaningful satellite imagery, suffering from severe artifacts and repeating patterns. In contrast, our proposed GeoCore-9B successfully retrieves geographic priors and synthesizes highly accurate terrains corresponding to the given coordinates.

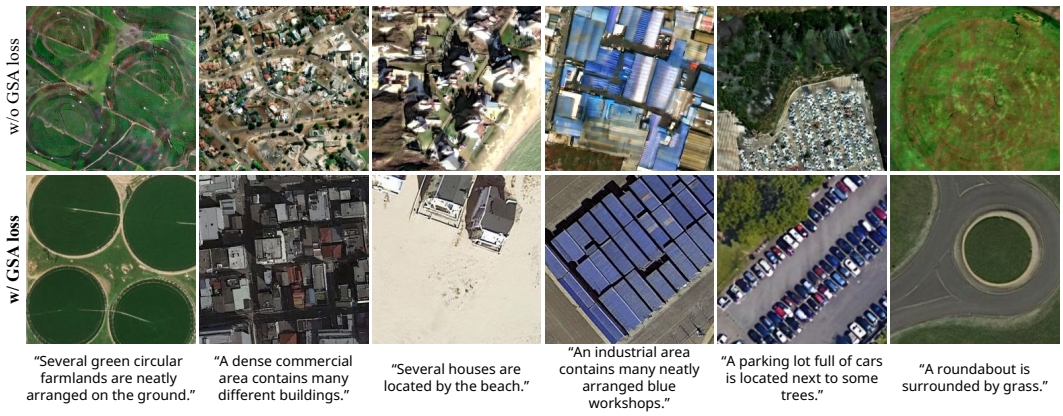

Figure 5: Ablation study on the Geospatial Semantic Alignment (GSA) loss. The inclusion of GSA loss (bottom row) improves structural fidelity and geometric consistency compared to the baseline without alignment (top row), which suffers from distorted boundaries and fragmented patterns.

boundaries, especially for structured scenes such as circular farmlands, industrial roofs, parking lots, and roundabouts. In contrast, GSA produces cleaner layouts and sharper object boundaries by aligning intermediate DiT features with satellite-specialist representations. Fig. 6 further shows that GSA consistently reduces FID on a 10K Git-10M subset across training iterations, indicating faster convergence and improved structural fidelity without additional inference cost.

4.5 RSICD Adaptation

We evaluate the lightweight text-to-image adaptation capabilities of our model on the RSICD [28] dataset. Following prior works [23, 41], we fine-tune LoRA adapters on the RSICD image-text training pairs. Since RSICD does not provide GSD values, latitudes, or longitudes, we employ learned null geospatial embeddings during both fine-tuning and inference. As shown in Table 1, GeoCore-9B significantly outperforms previous text-to-image methods on the RSICD benchmark across all evaluated metrics (Inception score, FID score, and CLIP score).

4.6 Practical and Challenging Downstream Tasks

Beyond text-to-image generation and controllability-oriented proxy tasks, we evaluate whether GeoCore-9B can be adapted to practical but challenging applications. We consider two image-

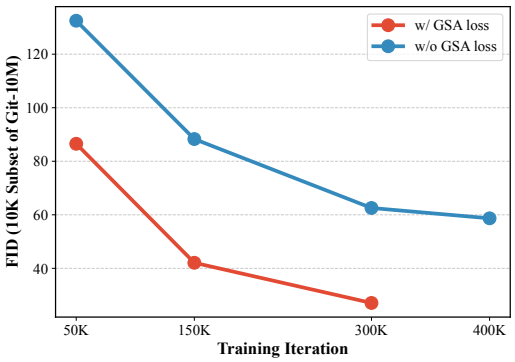

Figure 6: Ablation study on the Geospatial Semantic Alignment (GSA) loss. The inclusion of GSA loss improves structural fidelity.

Table 1: Comparison with previous text-to-image methods on the RSICD [28] dataset. **Bold** indicates the best result.

Method	IS ↑	FID ↓	CLIP ↑
Attn-GAN [45]	11.71	95.81	20.19
DAE-GAN [36]	7.71	93.15	19.69
StrucGAN [53]	5.84	–	–
DF-GAN [42]	9.51	109.41	19.76
Lafite [55]	10.70	74.11	22.52
DALL-E [34]	2.59	191.93	20.13
Txt2Img-MHN [46]	5.99	102.44	20.27
RSDiff [38]	7.22	66.49	–
CRS-Diff [41]	18.39	50.72	20.33
Text2Earth [23]	–	24.49	25.62
GeoCore-9B (Ours)	**22.15**	**18.82**	**27.15**

Table 2: Practical downstream adaptation results of GeoCore-9B. We evaluate cloud removal on Sen2-MTC [10] and SAR-to-optical cross-modal translation on QXS-SAROPT [11]. GeoCore-9B is adapted by simple fine-tuning and compared with task-specific or adapted baselines. **Bold** and underline indicate the best and second-best results, respectively.

(a) Cloud Removal

Methods	PSNR↑	SSIM↑	LPIPS↓
***Task-specific specialist** methods*			
McGAN [5]	17.448	0.513	0.447
Pix2Pix [12]	16.985	0.455	0.535
DSen2-CR [29]	16.827	0.534	0.446
STGAN [37]	18.152	0.587	0.513
CTGAN [10]	18.308	0.609	0.384
CR-TS-Net [4]	18.585	0.615	0.342
PMAA [57]	18.369	0.614	0.392
UnCRtainTS [3]	18.770	0.631	0.333
DDPM-CR [15]	18.742	0.614	0.329
DiffCR [58]	19.150	0.671	0.291
EMRDM [26]	20.067	0.709	**0.255**
Foundation model adaptation			
GeoCore-9B (Ours)	**20.809**	**0.799**	0.256

(b) SAR-to-Optical Image Translation

Methods	FID↓	LPIPS↓	HF-SCC↑	SSIM↑
***Task-specific or adapted** baselines*				
Pix2Pix [12]	196.89	0.454	0.0000	0.247
CycleGAN [56]	195.38	0.455	0.0001	0.251
SAR-SMTNet [49]	117.69	0.435	0.0003	0.260
CFCA-SET [19]	79.06	0.406	0.0006	0.273
BBDM [21]	65.15	0.522	0.0004	0.238
ControlNet [52]	22.39	0.434	0.0001	0.257
Uni-ControlNet [54]	22.48	0.437	0.0002	0.257
StegoGAN [44]	85.60	0.391	0.0019	0.280
DGDM [48]	147.23	0.634	0.0001	0.288
cBBDM [17]	69.47	0.420	0.0023	0.304
C-DiffSET [2]	18.15	**0.293**	0.0108	**0.372**
Foundation model adaptation				
GeoCore-9B (Ours)	**12.05**	0.377	**0.3360**	0.370

conditioned tasks: cloud removal and SAR-to-optical image translation. For both tasks, GeoCore-9B is fine-tuned with LoRA while the pre-trained backbone remains frozen.

Cloud removal. We evaluate GeoCore-9B on Sen2-MTC [10], where the model reconstructs cloud-free RGB images from cloudy observations (satellite RGB input images). Table 2(a) shows quantitative comparisons of GeoCore-9B with task-specific specialist methods for the cloud removal task. GeoCore-9B outperforms task-specific cloud removal methods in PSNR and SSIM, while achieving LPIPS comparable to the strongest specialist baseline [26, 58]. Fig. 7(a) shows that GeoCore-9B removes cloud contamination while preserving roads, and field boundaries.

SAR-to-optical image translation. We further evaluate GeoCore-9B on QXS-SAROPT [11] for the SAR-to-optical translation task in Table 2(b). GeoCore-9B achieves the best FID and HF-SCC, and remains comparable to the task-specific C-DiffSET baseline [2] in SSIM. Fig. 7(b) shows that GeoCore-9B generates EO images with clearer man-made structures and more faithful spatial layouts than recent translation baselines. These results demonstrate that GeoCore-9B can compete with, or outperform, task-specific models on such practical restoration and cross-modal translation tasks.

4.7 Limitations and discussions

GeoCore-9B leaves several directions for further extension. First, although we validate its transferability on practical tasks such as cloud removal and SAR-to-optical translation, broader evaluations

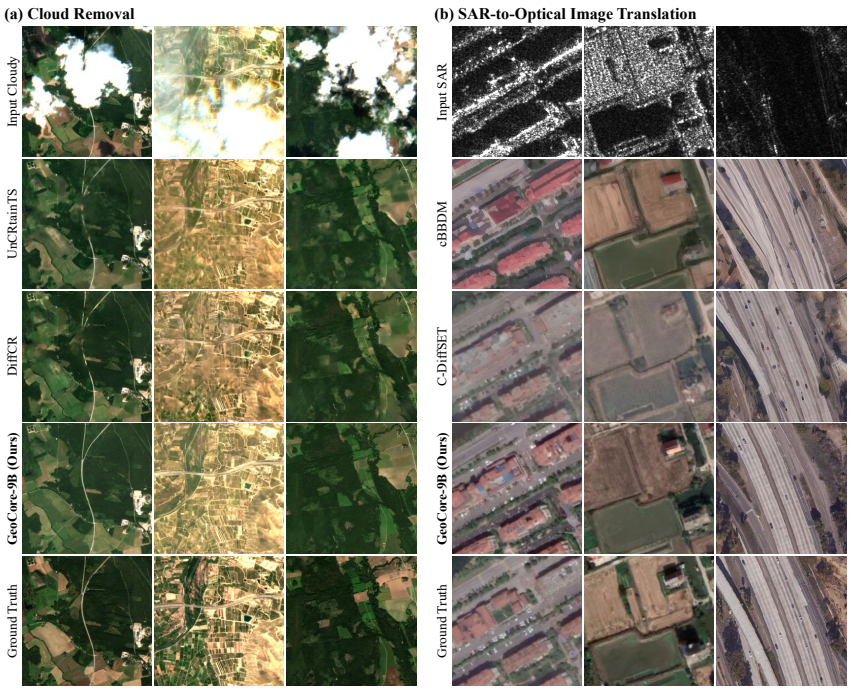

Figure 7: Qualitative comparison on practical downstream tasks. (a) Cloud removal: GeoCore-9B effectively removes heavy cloud contamination and reconstructs underlying structures (e.g., roads and field boundaries) much more faithfully than specialist baselines (UnCRtainTS [3] and DiffCR [58]). (b) SAR-to-optical image translation: GeoCore-9B translates noisy SAR inputs into realistic optical images, producing sharper man-made structures and more accurate spatial layouts compared to recent translation models (cBBDM [17] and C-DiffSET [2]).

remains as future work for additional real-world applications such as pan-sharpening, super-resolution, and segmentation-conditioned generation. Second, GeoCore-9B adopts a pre-trained VAE for efficient latent-space training for which future work will explore specialized latent learning for satellite imagery by jointly optimizing the VAE encoder with the DiT backbone, motivated by recent advances in representation and latent-space alignment [20, 47]. Finally, GeoCore-9B can be extended to handle multispectral satellite imagery based upon the large datasets with geo metadata and rich text prompts.

5 Conclusion

We presented GeoCore-9B, a 9-billion-parameters generative foundation model trained from the scratch for Earth Observation. Built upon a Flow Matching-based Diffusion Transformer, GeoCore-9B conditions generation on text and geospatial metadata, reducing reliance on natural image priors. To stabilize training at this large scale, we introduced a Geospatial Semantic Alignment loss, which distills structural Earth surface priors from a frozen DINOv3-Sat teacher network during training without adding inference overhead. Experiments show that GeoCore-9B achieves strong geo-aware generation capability and can be efficiently adapted to practical downstream tasks, including cloud removal and SAR-to-optical translation. These results suggest that large-scale generative pre-training on satellite RGB data provides a promising foundation for geo-aware remote sensing generation.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) under the Sejong Science Fellowship Program (RS-2026-25484549), for the project “Visualizing the Invisible Earth: A Reliability-Aware All-in-One SAR Analysis Framework with Foundation Models.”

A Broader Impacts

The development of GeoCore-9B presents both significant positive potential and dual-use risks. On the positive side, our generative foundation model can substantially advance Earth Observation (EO) applications by providing high-quality data augmentation for scarce regions, and by enhancing downstream tasks such as environmental monitoring, disaster response, and urban planning. Conversely, the ability to generate highly realistic, geo-aware synthetic satellite imagery introduces the risk of creating geographical deepfakes. If misused, such technology could be exploited to generate disinformation regarding geopolitical events, natural disasters, or environmental conditions. To mitigate these negative societal impacts, we emphasize the necessity of developing robust synthetic image detection frameworks tailored specifically for satellite imagery and advocate for the responsible deployment and disclosure of generative EO models.

B Validation of the Pre-trained VAE on EO Data

GeoCore-9B operates in the latent space of a frozen variational autoencoder (VAE) [18] originally optimized for natural images. The VAE denotes the complete encoder–decoder model; below, $\mathcal{E}$ and $\mathcal{D}$ denote its encoder and decoder, respectively. For an input $\mathbf{x}$, the reconstruction $\widehat{\mathbf{x}} = \mathcal{D}(\mathcal{E}(\mathbf{x}))$ sets an upper bound on input-faithful detail: structures discarded by $\mathcal{E}$ cannot be reliably recovered by the DiT. We therefore audit this bottleneck rather than assuming that a natural-image VAE is lossless on EO imagery.

Under the submitted frozen-VAE encode–decode protocol, a random set of 100,000 Git-10M [23] images gives 31.48 dB PSNR, 0.9310 SSIM, and 0.710 high-pass spatial correlation coefficient (HF-SCC). These averages indicate strong reconstruction fidelity and substantial preservation of high-frequency structure at 256×256 resolution. They do not, however, separately test the most extreme frequency bands or guarantee the preservation of small, sparse targets.

The positive HF-SCC values, ranging from 0.416 to 0.782 on the downstream domains, support the VAE's practical use for the evaluated 256×256 tasks, including replicated-SAR inputs. The lower Sen2-MTC values also expose a genuine domain-dependent limitation. After task-specific radiometric preprocessing, many Sen2-MTC pixels and local variations have low contrast; the natural-image VAE preserves the dominant coarse content but attenuates some weak local variations. This behavior is consistent with the lower HF-SCC, although that aggregate metric alone cannot establish the precise cause. We therefore claim task- and resolution-bounded suitability, not losslessness or guaranteed preservation of every EO microstructure.

C Additional Controlled Analyses

This section reports the controlled experiments completed after the original submission. We separate three questions: whether GSA helps at fixed scale, whether the fixed checkpoint responds to metadata interventions, and whether coordinate-only generations retrieve near-duplicates from the pre-training corpus. None of these experiments is presented as a full decomposition of model scale, data scale, and compute.

C.1 Matched 9B Ablation of GSA

We compare two DiT backbones trained from random initialization with the same 9B architecture, Git-10M data, 256×256 resolution, pre-training budget, and optimization protocol. Their downstream adaptation schedules are also identical; the controlled variable is only the GSA weight, $\mu = 0$ versus $\mu = 0.5$. The GSA teacher and projection head are discarded after pre-training and add no inference-time module or cost.

GSA improves every reported metric. In particular, it reduces FID by 9.61 points on RSICD and 7.87 points on QXS-SAROPT, while improving Sen2-MTC PSNR by 1.256 dB and SSIM by 0.116. This matched comparison isolates a benefit from GSA within the tested EO-trained 9B setting. It does not isolate the effects of overall model size, EO data, compute, or their interactions, and therefore does not explain the entire margin to external baselines.

Figure 8: Qualitative evaluation of VAE reconstruction on Earth Observation data. Top row: Original input satellite images from Git-10M. Bottom row: Reconstructed images using the frozen pre-trained VAE. Despite the domain shift from natural images, the VAE accurately recovers fine-grained textures, complex building structures, and intricate field patterns without any EO-specific fine-tuning.

Table 3: Frozen-VAE encode–decode fidelity on the pre-training and downstream domains. SAR intensities are replicated from one channel to three channels before VAE encoding.

Domain	n	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	**HF-SCC** ↑
Git-10M RGB	100,000	31.48	0.931	–	0.710
QXS-SAROPT optical	2,000	38.93	0.968	0.011	0.766
QXS-SAROPT SAR (1ch → 3ch)	2,000	28.77	0.932	0.023	0.782
Sen2-MTC cloudy	687	35.48	0.943	0.013	0.416
Sen2-MTC cloud-free	687	35.58	0.924	0.016	0.564

C.2 Fixed-Checkpoint Metadata Interventions

We conduct paired inference-time interventions on 1,000 metadata-parseable Git-10M samples using the submitted zero-shot checkpoint. For each sample, we fix the caption, noise seed, checkpoint, Euler sampler, 50 sampling steps, and CFG scale of 4.0. We change only the metadata: full metadata, a learned-null GSD, GSD shuffled from another sample, learned-null coordinates, or a jointly shuffled latitude–longitude pair.

To measure whether the generated images reflect these interventions, we train linear probes on frozen DINOv3-Sat ViT-L features from 20,000 real images and validate on a disjoint set of 5,000 real images. Before applying the probes to generated images, the nine-bin GSD probe reaches 0.769 validation accuracy (majority: 0.375), and the location probe reaches 0.585 over 81 eligible 15° regions (majority: 0.136). Location results below use the 957 generated samples retained by the eligible-region criterion.

Nulling GSD lowers GSD-bin accuracy from 0.555 to 0.293, and nulling coordinates lowers region accuracy from 0.485 to 0.175. The coordinate shuffle provides the clearest intervention result: generated outputs agree more with the supplied regions than with the original regions (0.366 versus 0.110). For the GSD shuffle, supplied-condition accuracy exceeds original-condition accuracy (0.376 versus 0.286), but 0.376 is essentially the 0.375 majority baseline; we therefore do not use this cell alone as evidence of fine-grained GSD following. FID is numerically worse under every intervention, but these values are only descriptive within-protocol checks. Cross-field changes further indicate that GSD and location are not perfectly disentangled. Overall, the experiment shows that metadata interventions affect the fixed model's outputs; it does not quantify how much of the external-model performance margin arises from metadata rather than scale, data, or GSA.

C.3 Full-Corpus Near-Duplicate Retrieval

We encode all 10,503,567 pre-training images and 500 text-free, coordinate-only generations with frozen DINOv3-Sat global features. Before examining the generated queries, we fix the similarity threshold at $s_{\mathrm{thr}} = 0.948$, the 95th percentile of nearest-*other*-image similarities from 1,000 real calibration queries with self-matches excluded (calibration median: 0.882). The generated-query

Table 4: Matched downstream comparison with and without GSA. All settings other than the pre-training GSA weight are held fixed. HF-SCC uses the corrected, baseline-consistent definition.

Task	**w/ GSA ($\mu = 0.5$)**	**w/o GSA ($\mu = 0$)**
RSICD text-to-image	**22.15 IS / 18.82 FID / 27.15 CLIP**	19.16 IS / 28.43 FID / 24.21 CLIP
QXS-SAROPT translation	**12.05 FID / 0.377 LPIPS / 0.0163 HF-SCC / 0.370 SSIM**	19.92 FID / 0.436 LPIPS / 0.0098 HF-SCC / 0.324 SSIM
Sen2-MTC cloud removal	**20.809 PSNR / 0.799 SSIM / 0.256 LPIPS**	19.553 PSNR / 0.683 SSIM / 0.284 LPIPS

Table 5: Paired metadata interventions. "Orig." and "suppl." score a shuffled-condition output against its original and newly supplied metadata, respectively. FID values support comparisons only within this protocol.

Condition	**FID ↓**	**GSD-bin acc. ↑**	**15°-region acc. ↑**
Full metadata	48.32	0.555	0.485
GSD null	54.96	0.293	0.460
GSD shuffled	49.46	0.286 (orig.) / 0.376 (suppl.)	0.424
Coordinates null	68.18	0.261	0.175
Coordinates shuffled	51.20	0.492	0.110 (orig.) / 0.366 (suppl.)

nearest-neighbor similarities have median 0.791, 95th percentile 0.871, and maximum 0.928; none exceeds the threshold (0/500). This finite global-feature test finds no near-duplicate under the stated protocol, but it cannot exclude localized, transformed, or other forms of memorization. Moreover, coordinate-only generation removes text but is not a complete text-null distributional ablation.

D Exploratory Frozen-Feature Transfer Probes

Motivated by diffusion-feature probing in SatDiFuser [14], we test whether task-relevant information is linearly accessible from the submitted 256×256 GeoCore-9B checkpoint. These experiments are representation probes, not full task-specific systems.

We freeze the VAE and DiT. To avoid conflict with the Flow Matching notation in the main paper, let $\mathbf{z}_{\text{clean}} = \mathcal{E}(\mathbf{x})$ and define the probe input at noise level τ as

$$\mathbf{z}_\tau = (1 - \tau)\mathbf{z}_{\text{clean}} + \tau\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{6}$$

We use empty text and null metadata, concatenate the 16×16 image-token features from DiT blocks 4, 8, 16, and 24 into a 16,384-dimensional feature, and train only one linear layer. For dense tasks, the token grid is bilinearly upsampled to 64×64. We test the three fixed noise levels $\tau \in \{0.25, 0.50, 0.75\}$ and use $\tau = 0.50$ as the main reference because it is the interpolation midpoint, not because it was tuned per task.

As a calibrated discriminative control, DINOv3-Sat ViT-L uses its standard clean-input last-four-layer features. The control follows the same images, splits, token grid, upsampling, linear-head design, and training schedule, although its feature width and extraction path differ from GeoCore-9B.

At $\tau = 0.50$, GeoCore-9B is within 0.5 top-1 percentage points of the DINOv3-Sat control on EuroSAT, reaches approximately 85% of its LoveDA mIoU, and exceeds it on BRIGHT (0.482 versus 0.442). Repeating the BRIGHT linear probe on the same frozen features gives 0.466 mIoU, still above the control. Classification varies by only 0.2 points across the tested noise levels, whereas both dense tasks perform best at the fixed midpoint.

These probes show linear accessibility of task-relevant information, but they are not comparisons with full-resolution specialist systems. DINOv3-Sat is also the GSA teacher, and we have not run a matched w/o-GSA feature probe; therefore, these results do not isolate how much of the transfer is caused by GSA. BRIGHT uses replicated-grayscale SAR only as an input, so this experiment also does not demonstrate SAR generation.

E Implementation Details

3D RoPE. To effectively model the joint sequence of text and latent image tokens, we employ a 3D Rotary Positional Embedding (3D RoPE) [40]. Specifically, we map each token into a unified

Table 6: Frozen-feature linear probes on EuroSAT [7], LoveDA [43], and BRIGHT [1]. The train/validation sizes are shown in parentheses.

Task (train/val)	**Metric**	$\tau = .25$	$\tau = .50$	$\tau = .75$	**DINOv3-Sat**
EuroSAT (12,960/3,240)	Top-1	97.3%	97.5%	97.3%	**98.0%**
LoveDA (3,000/1,200)	mIoU	0.328	0.382	0.282	**0.451**
BRIGHT (2,500/349 pairs)	mIoU	0.402	**0.482**	0.381	0.442

3D coordinate system (x, y, z). For textual tokens, the x-axis represents the 1D sequence index m, while for visual tokens, the (y, z) axes correspond to the 2D spatial grid coordinates (i, j). This 3D formulation allows the model to inherently reason about relative distances both within and across modalities, maintaining robust geospatial structural reasoning even under dynamic variations in image resolution and text length.

Classifier-Free Guidance. For Classifier-Free Guidance (CFG) [8], we implement an independent condition dropout strategy during training. While the text prompt is simply replaced by an empty string, masking continuous geographic metadata requires a more robust formulation. Thus, we introduce explicit learnable null embeddings e_r, e_ϕ, and e_λ to substitute the missing GSD, latitude, and longitude features, respectively. With a probability p_{cfg} (e.g., 0.1), all elements in the conditioning set $\mathcal{C} = \{d, r, \phi, \lambda\}$ are jointly replaced by their null counterparts to learn a fully unconditional prior. Otherwise, each condition is independently masked with its own dropout probability. This rigorous formulation ensures the model captures both the joint and marginal distributions of the geospatial and textual priors.

F More Results and Qualitative Diversity

In this section, we provide additional qualitative results to further demonstrate the generative capabilities, diversity, and downstream transferability of GeoCore-9B.

Text-conditioned generation. Figure 9 presents a broader set of text-conditioned generation examples. As shown, GeoCore-9B consistently synthesizes highly diverse and structurally realistic satellite imagery across various complex textual prompts. Compared to existing baseline models such as CRS-Diff [41] and Text2Earth [23], which frequently exhibit unnatural textures and structural artifacts, our model strictly maintains the authentic orthographic geometry of Earth observation data without losing fine-grained details.

GSD-conditioned generation. Figure 10, Figure 11, and Figure 12 provide further qualitative results demonstrating the scale-awareness of GeoCore-9B across varying ground sample distances (GSD). As the spatial resolution transitions from fine-grained (e.g., $1m$) to coarse-grained (e.g., $32m$), GeoCore-9B adaptively adjusts its visual granularity. It seamlessly shifts from synthesizing detailed individual objects to rendering broad, macroscopic land-cover patterns. Unlike baseline models that frequently struggle with scale inconsistency—producing unnaturally sized objects or repetitive textures at extreme resolutions—our model maintains strict physical and structural fidelity corresponding to the exact target GSD.

Cloud removal. Beyond zero-shot generation, we provide extended qualitative comparisons for practical downstream restoration tasks. Figure 13 illustrates additional results for the cloud removal task. Even under heavy and heterogeneous cloud coverage, GeoCore-9B accurately recovers the underlying geographic contexts, such as intricate road networks, detailed field boundaries, and diverse land-cover types. It notably produces much more faithful and sharper reconstructions than specialist models like UnCRtainTS [3] and DiffCR [58], which often yield blurry or semantically inconsistent regions.

SAR-to-optical image translation. Finally, Figure 14 showcases more examples of SAR-to-optical cross-modal translation. The inherently noisy and speckle-heavy nature of Synthetic Aperture Radar (SAR) imagery makes this task particularly challenging. Nevertheless, GeoCore-9B effectively translates these noisy inputs into clear, high-fidelity optical images. It excels at generating accurate spatial layouts and sharp man-made structures, demonstrating clear visual superiority over recent state-of-the-art translation models, including cBBDM [17] and C-DiffSET [2].

G Scope, Attribution, and Limitations

Contribution and attribution. GeoCore-9B uses a standard Flow Matching DiT backbone; we do not claim a new generic Flow Matching objective or transformer block, and numerical geospatial conditioning is not itself a new primitive. The system contribution is the EO-trained 9B generative backbone and its release. The specific design contributions are the integration of continuous GSD and coordinate conditioning and GSA as an EO-specialist representation-alignment instantiation. The matched experiment in Table 4 isolates GSA at fixed 9B architecture, data, and training scale, while the interventions in Table 5 establish fixed-model responsiveness to metadata. Neither experiment decomposes the effects of model size, EO data, compute, metadata, and GSA relative to external systems. A full factorial study varying these factors at 9B scale remains outside the present scope. Git-10M is a public dataset, and we do not claim its construction as a contribution.

CRS-Diff [41] is best interpreted as a controllable EO generator, whereas Text2Earth [23] is the more direct EO text-to-image comparator. Closed general-purpose generators do not expose weights, training data, or fine-tuning access and therefore provide, at most, uncontrolled qualitative references rather than matched quantitative evidence.

Residual natural-image components and the VAE ceiling. The 9B generative DiT backbone is initialized and trained from scratch on EO data, but the complete system retains an off-the-shelf natural-image VAE and pretrained text encoders. Thus, the model avoids initialization of its DiT from a natural-image diffusion backbone; it is not entirely free of natural-image priors. The VAE is also an information bottleneck whose reconstruction quality upper-bounds input-faithful detail. The audits in Sec. B support practical use for the current 256×256 tasks but do not guarantee preservation of every tiny target or high-frequency structure. A promising direction is EO-specific end-to-end adaptation of the VAE and DiT, following representation-aligned joint tuning such as REPA-E [20], while retaining reconstruction and EO-specific high-frequency preservation objectives.

RGB output and sensor scope. GeoCore-9B currently generates RGB optical imagery. In SAR-to-optical translation, SAR is only an input condition and the target remains RGB optical; the current model does not generate SAR or multispectral imagery. DINOv3-Sat never directly processes SAR in the submitted system: GSA is used only during RGB EO pre-training to supervise intermediate optical-target representations, and the teacher and projection head are removed before downstream adaptation and inference. Consequently, SAR measurements are not directly forced to match the RGB teacher space. Nevertheless, GSA can leave an optical prior in the learned backbone weights. The improved QXS-SAROPT results establish compatibility with the tested SAR-conditioned, RGB-output setting, not sensor-universal transfer. SARMAE [24] provides complementary evidence that optical DINOv3 supervision can benefit SAR representation learning, but broader output modalities would still require modality-appropriate latent encoders and modality-specific or multimodal teachers.

Discriminative transfer. The exploratory probes in Sec. D show that task-relevant information is linearly accessible from frozen GeoCore-9B features. They use a low-resolution token grid and a single linear head, rather than full task-specific systems, and do not include a matched w/o-GSA probe. They therefore neither establish discriminative state of the art nor attribute the observed transfer specifically to GSA.

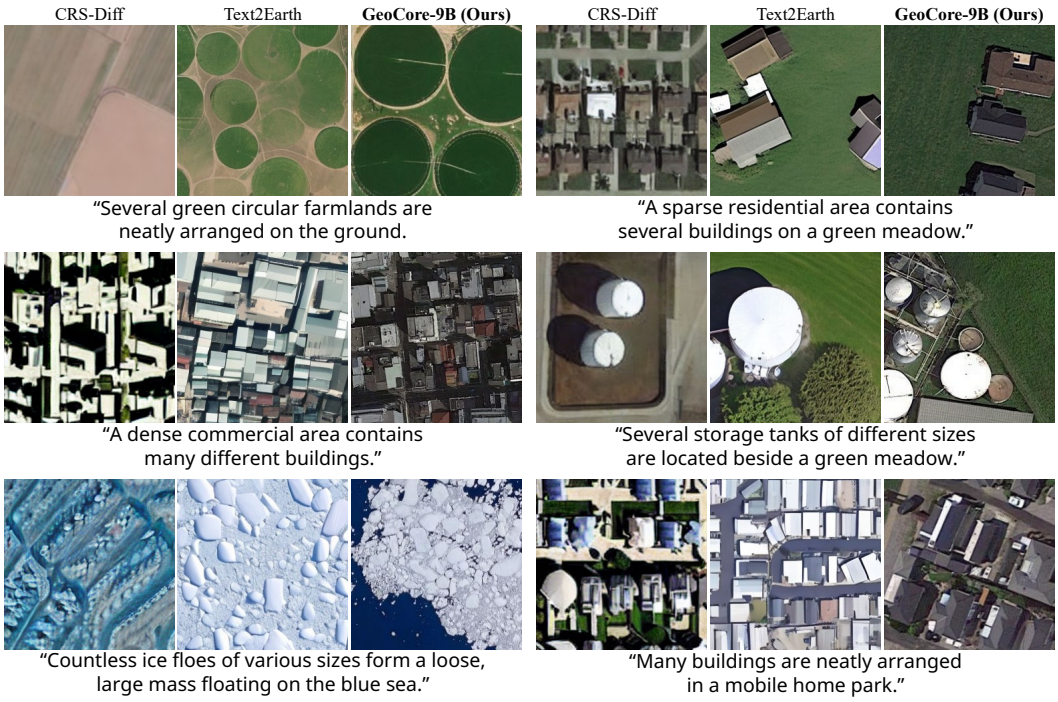

Figure 9: Qualitative comparison of text-conditioned generation. Given various text descriptions, GeoCore-9B synthesizes highly realistic and structurally accurate satellite imagery, outperforming baselines (CRS-Diff [41] and Text2Earth [23]) which often suffer from severe artifacts.

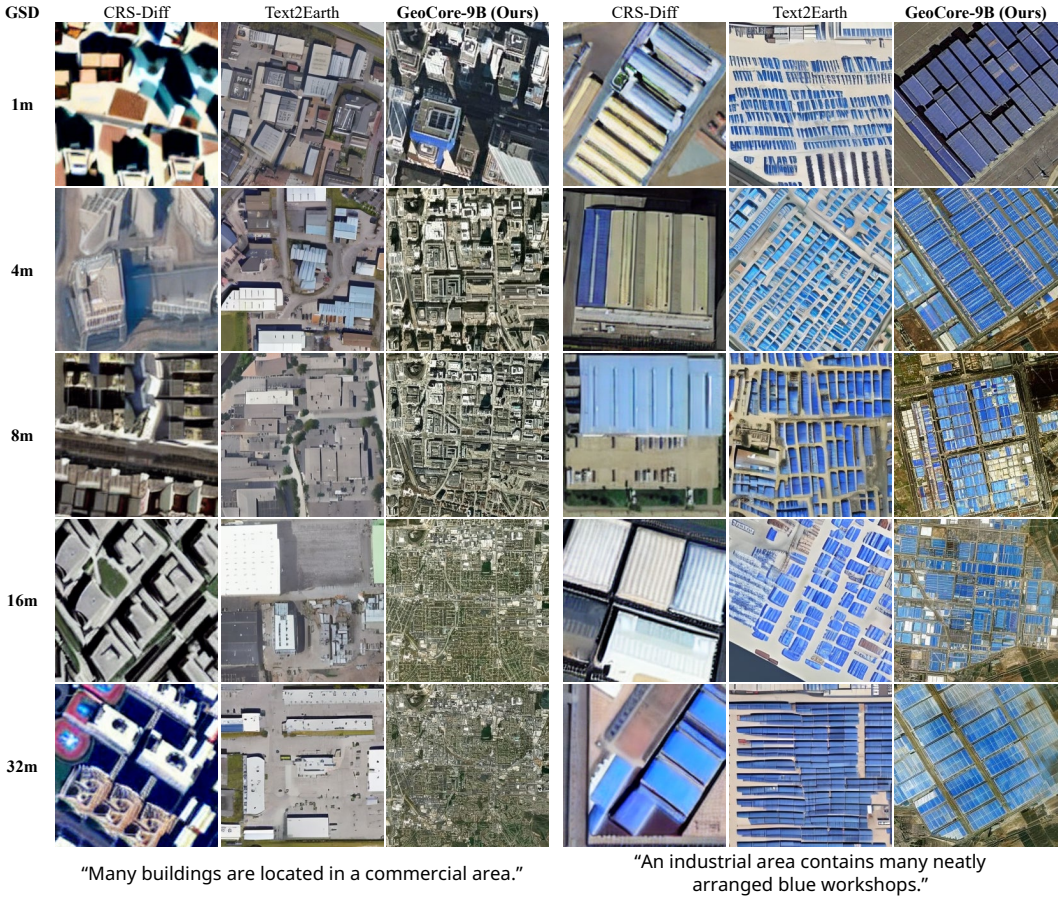

Figure 10: Qualitative comparison across varying ground sample distances (GSD). GeoCore-9B adaptively adjusts visual granularity from fine structural details ($1m$) to broad land-cover patterns ($32m$), demonstrating superior scale-awareness compared to CRS-Diff [41] and Text2Earth [23] which struggle with unnatural textures and scale inconsistency.

Figure 11: Qualitative comparison across varying ground sample distances (GSD). GeoCore-9B adaptively adjusts visual granularity from fine structural details ($1m$) to broad land-cover patterns ($32m$), demonstrating superior scale-awareness compared to CRS-Diff [41] and Text2Earth [23] which struggle with unnatural textures and scale inconsistency.

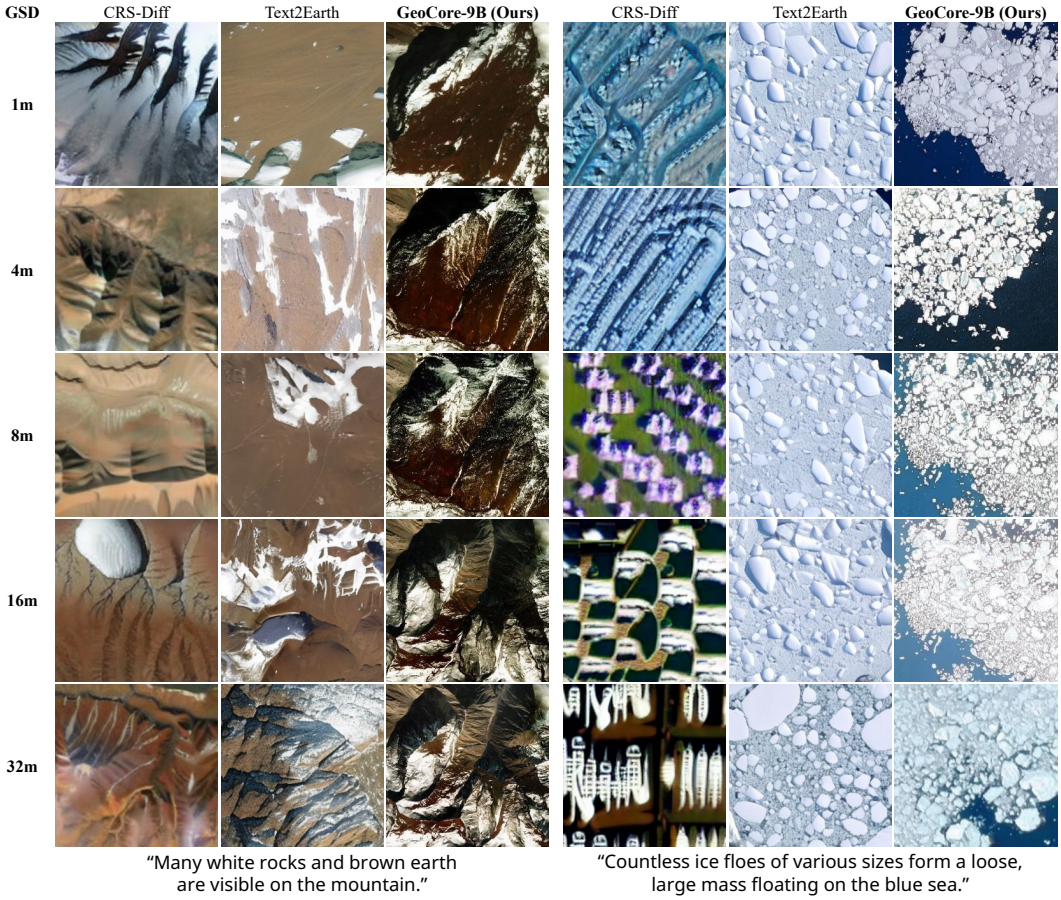

Figure 12: Qualitative comparison across varying ground sample distances (GSD). GeoCore-9B adaptively adjusts visual granularity from fine structural details ($1m$) to broad land-cover patterns ($32m$), demonstrating superior scale-awareness compared to CRS-Diff [41] and Text2Earth [23] which struggle with unnatural textures and scale inconsistency.

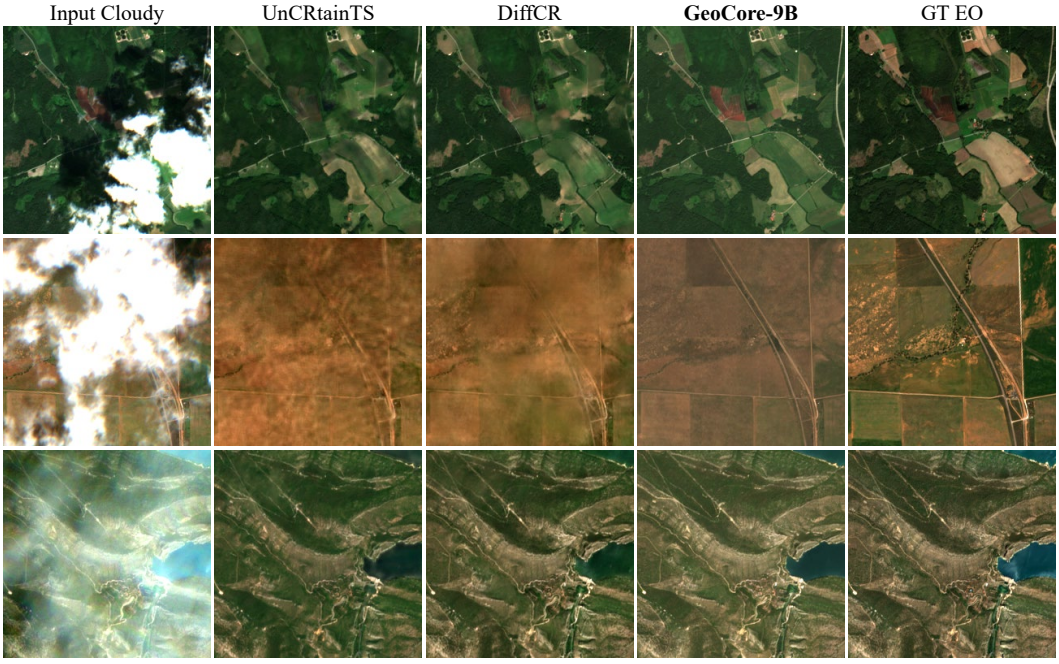

Figure 13: Qualitative comparison on practical downstream tasks (cloud removal). GeoCore-9B effectively removes heavy cloud contamination and reconstructs underlying structures (e.g., roads and field boundaries) much more faithfully than specialist baselines (UnCRtainTS [3] and DiffCR [58]).

Figure 14: Qualitative comparison on practical downstream tasks (SAR-to-optical image translation). GeoCore-9B translates noisy SAR inputs into realistic optical images, producing sharper man-made structures and more accurate spatial layouts compared to recent translation models (cBBDM [17] and C-DiffSET [2]).

References

[1] Hongruixuan Chen, Jian Song, Olivier Dietrich, Clifford Broni-Bediako, Weihao Xuan, Junjue Wang, Xinlei Shao, Yimin Wei, Junshi Xia, Cuiling Lan, Konrad Schindler, and Naoto Yokoya. Bright: A globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response. *Earth System Science Data*, 17(11):6217–6253, 2025.

[2] Jeonghyeok Do, Jaehyup Lee, and Munchurl Kim. C-diffset: Leveraging latent diffusion for sar-to-eo image translation with confidence-guided reliable object generation. *arXiv preprint arXiv:2411.10788*, 2024.

[3] Patrick Ebel, Vivien Sainte Fare Garnot, Michael Schmitt, Jan Dirk Wegner, and Xiao Xiang Zhu. Uncrtaints: Uncertainty quantification for cloud removal in optical satellite time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2086–2096, 2023.

[4] Patrick Ebel, Yajin Xu, Michael Schmitt, and Xiao Xiang Zhu. Sen12ms-cr-ts: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[5] Kenji Enomoto, Ken Sakurada, Weimin Wang, Hiroshi Fukui, Masashi Matsuoka, Ryosuke Nakamura, and Nobuo Kawaguchi. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 48–56, 2017.

[6] 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 *Forty-first international conference on machine learning*, 2024.

[7] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[10] Gi-Luen Huang and Pei-Yuan Wu. Ctgan: Cloud transformer generative adversarial network. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 511–515. IEEE, 2022.

[11] Meiyu Huang, Yao Xu, Lixin Qian, Weili Shi, Yaqin Zhang, Wei Bao, Nan Wang, Xuejiao Liu, and Xueshuang Xiang. The qxs-saropt dataset for deep learning in sar-optical data fusion. *arXiv preprint arXiv:2103.08259*, 2021.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[13] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7383–7394, 2025.

[14] Yuru Jia, Valerio Marsocci, Ziyang Gong, Xue Yang, Maarten Vergauwen, and Andrea Nascetti. Can generative geospatial diffusion models excel as discriminative geospatial foundation models? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[15] Ran Jing, Fuzhou Duan, Fengxian Lu, Miao Zhang, and Wenji Zhao. Denoising diffusion probabilistic feature-based network for cloud removal in sentinel-2 imagery. *Remote Sensing*, 15(9):2217, 2023.

[16] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B. Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] Seon-Hoon Kim and Daewon Chung. Conditional brownian bridge diffusion model for vhr sar to optical image translation. *IEEE Geoscience and Remote Sensing Letters*, 2025.

[18] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

[19] Jaehyup Lee, Hyebin Cho, Doochun Seo, Hyun-Ho Kim, Jaeheon Jeong, and Munchurl Kim. Cfca-set: Coarse-to-fine context-aware sar-to-eo translation with auxiliary learning of sar-to-nir translation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023.

[20] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18262–18272, 2025.

[21] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023.

[22] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[23] 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*, 2025.

[24] Danxu Liu, Di Wang, Hebaixu Wang, Haoyang Chen, Wentao Jiang, Yilin Cheng, Haonan Guo, Wei Cui, and Jing Zhang. Sarmae: Masked autoencoder for sar representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[25] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[26] Yi Liu, Wengen Li, Jihong Guan, Shuigeng Zhou, and Yichao Zhang. Effective cloud removal for remote sensing images by an improved mean-reverting denoising model with elucidated design space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17851–17861, 2025.

[27] Yidan Liu, Jun Yue, Shaobo Xia, Pedram Ghamisi, Weiying Xie, and Leyuan Fang. Diffusion models meet remote sensing: Principles, methods, and perspectives. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–22, 2024.

[28] 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.

[29] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346, 2020.

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[32] 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.

[33] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pages 1–16. IEEE, 2020.

[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[36] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13960–13969, 2021.

[37] Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon. Cloud removal from satellite images using spatiotemporal generator networks. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1796–1805, 2020.

[38] Ahmad Sebaq and Mohamed ElHelw. Rsdiff: Remote sensing image generation from text using diffusion model. *Neural Computing and Applications*, 36(36):23103–23111, 2024.

[39] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[40] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[41] Datao Tang, Xiangyong Cao, Xingsong Hou, Zhongyuan Jiang, Junmin Liu, and Deyu Meng. Crs-diff: Controllable remote sensing image generation with diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[42] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16515–16525, 2022.

[43] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, volume 1, 2021.

[44] Sidi Wu, Yizi Chen, Samuel Mermet, Lorenz Hurni, Konrad Schindler, Nicolas Gonthier, and Loic Landrieu. Stegogan: Leveraging steganography for non-bijective image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7922–7931, 2024.

[45] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

[46] 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, 2023.

[47] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 2025.

[48] Donggeun Yoon, Minseok Seo, Doyi Kim, Yeji Choi, and Donghyeon Cho. Deterministic guidance diffusion model for probabilistic weather forecasting. *arXiv preprint arXiv:2312.02819*, 2023.

[49] Geunhyuk Youk and Munchurl Kim. Transformer-based synthetic-to-measured sar image translation via learning of representational features. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023.

[50] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

[51] 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.

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[53] Rui Zhao and Zhenwei Shi. Text-to-remote-sensing-image generation with structured generative adversarial networks. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[54] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in neural information processing systems*, 36:11127–11150, 2023.

[55] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17907–17917, 2022.

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[57] Xuechao Zou, Kai Li, Junliang Xing, Pin Tao, and Yachao Cui. Pmaa: A progressive multi-scale attention autoencoder model for high-performance cloud removal from multi-temporal satellite imagery. *arXiv preprint arXiv:2303.16565*, 2023.

[58] Xuechao Zou, Kai Li, Junliang Xing, Yu Zhang, Shiying Wang, Lei Jin, and Pin Tao. Diffcr: A fast conditional diffusion framework for cloud removal from optical satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.