GeoDiT: A Diffusion-based Vision-Language Model for Geospatial Understanding

Jiaqi Liu, Haoran Liu, Lang Sun
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University
Changchun, Jilin 130012, China
{liujq21,lhr24,sunlang24}@mails.jlu.edu.cn

Ronghao Fu*, Bo Yang*
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University
Changchun, Jilin 130012, China
{furh,ybo}@jlu.edu.cn

Abstract

Autoregressive models are structurally misaligned with the inherently parallel nature of geospatial understanding, forcing a rigid sequential narrative onto scenes and fundamentally hindering the generation of structured and coherent outputs. We challenge this paradigm by reframing geospatial generation as a parallel refinement process, enabling a holistic, coarse-to-fine synthesis that resolves all semantic elements simultaneously. To operationalize this, we introduce GeoDiT, the first diffusion-based vision-language model tailored for the geospatial domain. Extensive experiments demonstrate that GeoDiT establishes a new state-of-the-art on benchmarks requiring structured, object-centric outputs. It achieves significant gains in image captioning, visual grounding, and multi-object detection, precisely the tasks where autoregressive models falter. Our work validates that aligning the generative process with the data's intrinsic structure is key to unlocking superior performance in complex geospatial analysis.

1. Introduction

Automating the interpretation of the vast and ever-growing stream of geosptial data constitutes a pivotal challenge in the remote sensing domain. From monitoring deforestation for climate security [27, 39] to tracking infrastructure development for urban dynamics [3, 11], the ability to rapidly transform raw satellite imagery into actionable insights is critical. In response, adapting large-scale vision-language models (VLMs) on Earth observation data has emerged as the dominant and most promising paradigm.

This paradigm has largely evolved along two architectural lines. Initial efforts centered on aligning visual and

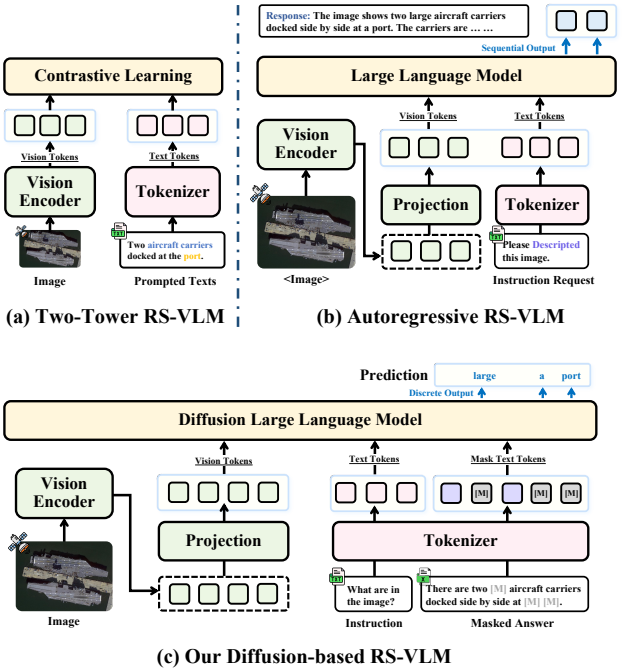


Figure 1. Conceptual comparison of text generation paradigms for geospatial understanding. (a) A two-tower model based on contrastive learning. (b) An autoregressive model generating text sequentially. (c) Our proposed diffusion-based model generating text in parallel via iterative mask prediction.

textual representations within a shared embedding space through two-tower contrastive learning (Figure 1(a)). While effective for retrieval-style tasks, these models [20, 22, 48] lack the capacity for nuanced, generative interpretation. More recently, the frontier has shifted to autoregressive VLMs, which fuse visual features directly into a Large Language Model backbone (Figure 1(b)). This architectural shift has rapidly consolidated into the prevailing methodology for generative geospatial intelligence, exemplified by a series of recent works such as GeoChat [17], VHM [29],

*Corresponding authors.

and EarthDial [32].

Undeniably, the success of this autoregressive paradigm has unlocked unprecedented capabilities in geospatial intelligence on a series of tasks such as scene classification (SC), visual question answering (VQA), visual grounding (VG). However, their celebrated success in these areas, which typically demand a single, focused output, masks a crucial architectural limitation. Specifically, their strictly sequential nature is fundamentally incompatible with a coarse-to-fine synthesis, a process that first establishes a scene's global composition (such as its principal subjects for a description or the number and placement of objects for detection) before committing to the fine-grained generation of specific words or coordinates. This linear, token-by-token commitment stands in stark contrast to human cognition, where a global gist of the scene precedes the articulation of details.

This core incompatibility, therefore, materializes as observable, systematic failures in tasks requiring structured outputs. In comprehensive scene description, the model's generative focus becomes prematurely anchored to the first salient entity, causing it to expend its descriptive budget on initial observations while struggling to cohesively integrate other, spatially-distant concepts into a balanced narrative. In multi-object detection, the same path-dependency creates a contextual feedback loop: the generation of one bounding box pathologically influences the prediction of the next, leading to the redundant detection of a single object, in place of a systematic canvass for other distinct entities. Ultimately, these failures reveal a shared root cause: the inherent inability of a sequential process to formulate a globally consistent understanding before committing to local, token-by-token generation.

Overcoming this limitation requires a new generative paradigm with an inherently global and parallel structure. The principles of denoising diffusion models [16, 19, 40, 42], which have recently catalyzed breakthroughs in general-purpose visual and language generation, offer a compelling and structurally-aligned solution to this challenge. Unlike autoregressive models, which are constrained to a rigid, unidirectional generation path, diffusion models frame generation as a parallel, iterative refinement of a complete output structure. As illustrated in Figure 1(c), they begin with a noisy, unstructured canvas and progressively denoise it, allowing all semantic units, be they words or coordinates, to be resolved simultaneously and co-dependently, thus building coherence at a global level from the outset.

Inspired by this structural alignment, we are the first to introduce this emerging paradigm to the geospatial domain, reframing the task of complex geospatial description as one of multimodally-conditioned text denoising. To operationalize this new paradigm, we present GeoDiT, a Diffusion Transformer-based vision-language model tailored for geospatial understanding. Through extensive experiments, we demonstrate that GeoDiT establishes new state-of-the-art results and validates the superiority of a structurally-aligned generative process. The principal contributions of this work are threefold:

- We challenge the prevailing autoregressive paradigm by reframing geospatial language generation as a problem of resolving a parallel, spatially-grounded semantic field.
- We operationalize this new paradigm in GeoDiT, a diffusion transformer-based model, providing the first systematic validation that iterative parallel decoding is a structurally superior approach for generating spatially-grounded semantic descriptions in remote sensing.
- We provide extensive empirical evidence on a suite of challenging benchmarks, demonstrating that the performance gains achieved by GeoDiT are a direct result of its structural alignment with the data, thereby validating the superiority of the proposed generative paradigm.

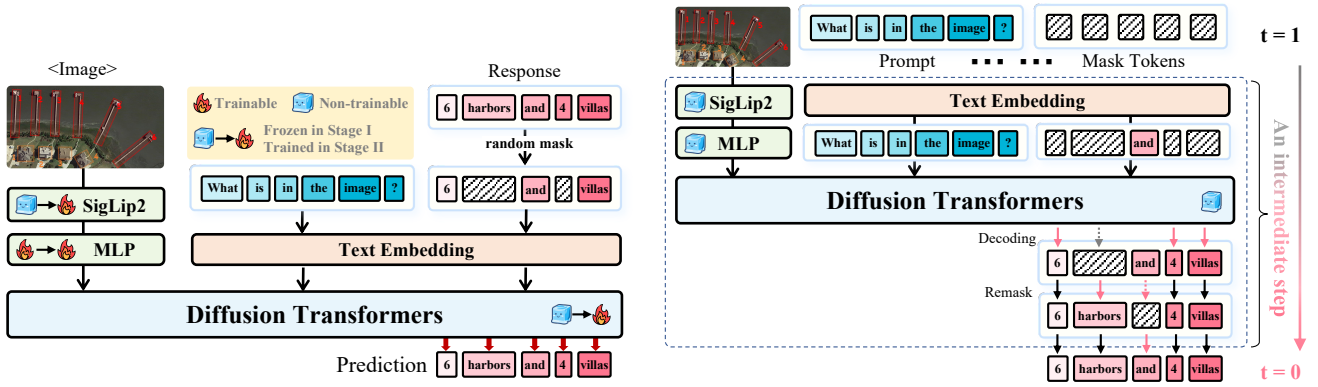
2. Related Work

2.1. Autoregressive RS Vision-Language Models

Recent advancements in Remote-sensing Vision-Language Models (RS-VLMs) have shown remarkable potential. Pioneering works such as GeoChat [17], VHM [29], and EarthDial [32] have successfully tackled tasks from scene parsing to VQA. However, despite these architectural and data-centric innovations, they are all fundamentally built upon an autoregressive framework. This shared foundation, while powerful for general narrative tasks, imposes a rigid sequential structure on their output. Consequently, they all inherit the same structural limitation when faced with the spatially independent and non-narrative nature of geospatial intelligence—a core problem our work addresses from an entirely different paradigm.

2.2. Diffusion-based Text Generation

Concurrent to the advancements in autoregressive models, a distinct, non-autoregressive (NAR) generation paradigm has emerged, recently culminating in powerful diffusion-based frameworks. Rather than generating through a sequential, token-at-a-time process, these models perform generation via a parallel, iterative refinement process, typically starting from a fully masked sequence and iteratively predicting the original content over discrete timesteps. This parallel refinement approach has been successfully operationalized in end-to-end multimodal systems such as LLaDA [42], LaViDA [15], and MMaDA [40], which have demonstrated significant potential for unified multimodal understanding on general-domain tasks. However, these powerful NAR techniques have been predominantly developed for and evaluated on narrative-style text. Their systematic architectural adaptation for the unique structural challenges of multimodal geospatial understanding remains a



(a) Training Model Architecture **(b) Inference Model Architecture**

Figure 2. Overview of the GeoDiT framework, illustrating the **(a)** training and **(b)** inference procedures. **(a)** During training, GeoDiT is optimized with a mask-and-predict objective. The model learns to reconstruct the original text from a randomly masked version, conditioned on a prompt and the visual features from a SigLIP-2 encoder. This follows a two-stage strategy: initial vision-language alignment by training only the MLP projector (Stage I), followed by end-to-end instruction tuning of the entire model (Stage II). **(b)** Inference is a non-autoregressive, iterative refinement process. Starting from a fully masked template, the model repeatedly predicts the full sequence and then applies a low-confidence remasking strategy. High-confidence tokens are preserved while uncertain ones are re-masked for the subsequent refinement step, a process that continues for a fixed number of iterations to produce the final output.

critical but unexplored research gap which we aim to solve.

3. Methodology

Our proposed framework, GeoDiT, introduces a diffusion-based paradigm to generate structured textual descriptions of remote sensing imagery. The complete architecture, illustrated in Figure 2, comprises a visual conditioning backbone and a core generative module, the Diffusion Transformer. In essence, the visual backbone first processes an input image to produce a set of conditioning vectors that encapsulate the image's geospatial context. These vectors are then utilized by the Diffusion Transformer, which performs a non-autoregressive, iterative decoding process to synthesize the final output. We detail the architecture of these components first, followed by a description of the training and inference procedures.

3.1. Model Architecture

The architecture of GeoDiT consists of two main modules: a visual backbone that provides geospatial context, and a generative core that synthesizes the textual output.

Visual Backbone. GeoDiT uses a pre-trained SigLIP-2 (ViT-SO400M) model [35] as its visual backbone to convert an input image $I \in \mathbb{R}^{H\times W\times 3}$ into a sequence of patch embeddings into a sequence of N visual patch embeddings:

$$Z_v = \text{Encoder}_{\text{ViT}}(I), \quad \text{where } Z_v \in \mathbb{R}^{N\times D_v}. \tag{1}$$

To align these visual features with the hidden dimension of our generative model, a lightweight Multi-Layer Perceptron (MLP) then projects Z_v to produce the final visual condition vectors:

$$C_v = \text{MLP}(Z_v), \quad \text{where } C_v \in \mathbb{R}^{N\times d}. \tag{2}$$

Here, D_v is the hidden dimension of the vision encoder, and d is the hidden dimension of our core Diffusion Transformer. The sequence C_v serves as the primary geospatial context for the generative process.

Generative Core: A Modality-Adapted DiT. At the heart of GeoDiT lies a generative module designed to synthesize structured text. We adopt the core architectural principle of Diffusion Transformers (DiT) [13]: leveraging a standard Transformer as a powerful, scalable backbone for a diffusion-based generative process. However, a fundamental adaptation is required to accommodate our target modality. While the original DiT operates on continuous latent variables via Gaussian diffusion, our objective is to generate discrete text tokens. We therefore reformulate the generative core to operate within a discrete, mask-and-predict diffusion framework.

Our generative core is a bidirectional Transformer, denoted as p_θ, which learns to reverse a forward masking process. Architecturally, its primary function is to predict the original identities of masked tokens in a sequence, conditioned on both unmasked textual context and the visual condition vectors C_v. At each reverse diffusion step t, the input to the Transformer is a hybrid sequence X_t formed by concatenating the visual vectors with the embeddings of the partially masked text sequence T_t:

$$X_t = \text{concat}\left(C_v, \mathbf{E}(T_t)\right), \tag{3}$$

where $\mathbf{E}$ is the text token embedding layer. This sequence is then processed by the stack of L Transformer blocks:

$$H_t = \text{Transformer}_\theta(X_t). \tag{4}$$

From the full sequence of output hidden states H_t, we select only those corresponding to the text tokens, denoted as H_t^{text}. These are then projected to produce a probability distribution over the entire vocabulary V for each masked position:

$$p_\theta(T_0|T_t, C_v) = \text{softmax}(\mathbf{W}_p H_t^{\text{text}} + \mathbf{b}_p), \tag{5}$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ are the parameters of the final linear projection layer.

In operationalizing this architecture, we identified the recent LLaDA-8B model [28] as a particularly suitable foundation. While not explicitly described as a Diffusion Transformer in its original presentation, its core mechanism—a bidirectional Transformer optimized for iterative mask-and-predict tasks—is fundamentally aligned with the discrete diffusion principles we require. This fundamental alignment establishes LLaDA-8B as a direct, pre-trained implementation of our proposed paradigm, whose bidirectional and iterative nature we have argued is essential for resolving the parallel semantics of geospatial data. Therefore, we adopt this validated architecture for GeoDiT, initializing our generative core with the publicly available LLaDA-8B weights to leverage its powerful foundational capabilities. Consequently, the primary novelty of our work lies not in redesigning this foundational architecture, but in the systematic methodology we developed to ground its powerful generative capabilities in the unique, non-narrative semantics of the geospatial domain.

3.2. Training Strategy

The training of GeoDiT is guided by a unified training objective, applied across a systematic, two-stage process. This approach is designed to first establish a fundamental vision-language alignment and then cultivate advanced instruction-following capabilities for the remote sensing domain.

Training Objective. The core objective is to train the parameters θ of the generative Transformer, p_θ, to reverse a discrete diffusion process. This process is formulated as a mask-and-predict task. Given a ground-truth text description $T_0 = (T_0^1, \ldots, T_0^L)$ and its corresponding image I, the training proceeds as follows. First, a random timestep $t \sim U[0, 1]$ is sampled. A forward process q then generates a corrupted version T_t by replacing each token T_0^i with a special '[M]' token, independently and with probability t.

The model p_θ is then optimized to predict the original tokens T_0 given the corrupted sequence T_t and the visual context from image I. This is achieved by minimizing the following loss function, $\mathcal{L}(\theta)$, which is an upper bound on the negative log-likelihood of the data:

$$\mathcal{L}(\theta) = \mathbb{E}_{(I,T_0)\sim\mathcal{D}} \left[-\sum_{i=1}^{L} \mathbb{1}[T_t^i = [\text{M}]] \log p_\theta(T_0^i|T_t, I) \right], \tag{6}$$

where $\mathcal{D}$ is the training dataset, $\mathbb{1}[\cdot]$ is the indicator function which activates the loss only for masked positions, and $p_\theta(T_0^i|T_t, I)$ is the probability assigned by the model to the correct token T_0^i at position i. This objective effectively

Task	Data	Size	Type
Image Captioning	RSICD [26]	24,333	optical
	UCM-Captions [5]	9,986	optical
	RSITMD [43]	20,096	optical
	Sydney-Captions [30]	2,837	optical
	NWPU-Captions [4]	141,631	optical
VQA	FloodNet [31]	4,511	optical
	RSVQA_LR [24]	67,228	optical
	RSIVQA [49]	68,625	optical
	CRSVQA [45]	900	optical
Classification	NWPU-RESISC45 [5]	6,300	optical
	EuroSAT [14]	5,400	optical
	UCMerced-Landuse [41]	420	optical
	WHU-RS19 [6]	196	optical
	RSSCN7 [52]	560	optical
	DSCR [8]	11,951	optical
	FGSCR-42 [7]	3,878	optical
Detection	DOTA [9]	163,486	optical
	DIOR [12]	58,200	optical
	FAIR1M [33]	40,466	optical
	NWPUVHR10 [18]	3,190	optical
	HRRSD [47]	23,044	optical
	RSOD [25]	747	optical
	UCAS-AOD [51]	1,510	optical
Visual Grounding	DIOR-RSVG [44]	30,820	optical
Region-level Captioning	DIOR-RSVG [44]	30,820	optical

Table 1. Details on the datasets collected from the MMRS-1M.

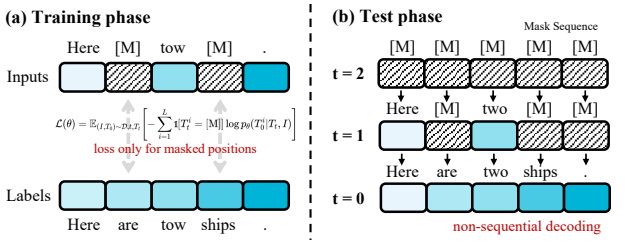


Figure 3. Illustration of the core mask-and-predict mechanism. (a) Training phase: The model is optimized to reconstruct the original text Labels from a randomly masked version Inputs. The loss is calculated only on the tokens that were masked. (b) Test phase: Generation is a non-sequential, iterative process. It starts from a fully masked sequence ($t = 2$) and progressively predicts the full sentence over a series of refinement steps ($t = 1, t = 0$).

trains the model to denoise the text by filling in the masked tokens based on multimodal context, as visually summarized in Figure 3(a).

Stage I: Vision-Language Alignment. We freeze the vision encoder and generative Transformer, training only the lightweight MLP projector to align the visual and text embedding spaces by optimizing the objective function in Eq. (6). We utilize SkyScript dataset [36] in this stage. This step effectively teaches the projector to map visual concepts into a representation that the generative core can readily understand, without the computational expense of updating the large backbone models.

Stage II: Full Instruction Tuning. We unfreeze all model components and perform end-to-end training on the optical subset of the MMRS-1M dataset shown in Table 1, created by amalgamating 34 remote sensing datasets into a unified instruction-following format. By minimizing the loss $\mathcal{L}(\theta)$ from Eq. (6), this end-to-end tuning allows all components to co-adapt, specializing the model for the nuanced task of remote sensing interpretation.

3.3. Inference Procedure

GeoDiT employs a non-autoregressive, iterative process that synthesizes the entire textual description by progressively denoising a masked sequence into the final response. This approach, illustrated in Figure 3(b), directly mirrors the reverse of the diffusion process learned during training.

Generation begins with a fully masked text template $T_{t=N} = \{[M], \ldots, [M]\}$ of a predetermined length L, corresponding to the initial timestep $t_N = 1$. The model then iteratively refines this template over a series of N discrete timesteps $(t_N, t_{N-1}, \ldots, t_1)$, where $t_1 \approx 0$. At each iteration k (from timestep t_k to t_{k-1}), the generative core p_θ first produces a complete prediction of the final text, denoted $\hat{T}_0$, by taking the most likely token at each position from the model's output distribution:

$$\hat{T}_0 = \underset{T'_0}{\text{argmax}}\, p_\theta(T'_0 | T_{t_k}, C_v), \tag{7}$$

where T_{t_k} is the masked text sequence at the current step.

This complete prediction, $\hat{T}_0$, serves as an intermediate estimate. To construct the input for the next, less-masked iteration, $T_{t_{k-1}}$, a scheduled remasking is applied. A key component of this process is the remasking strategy, which determines which tokens to reveal and which to refine. Following the effective approach in discrete diffusion models [28], we employ a **low-confidence remasking** strategy. Based on the confidence scores from the output probabilities $p_\theta(\cdot | T_{t_k}, C_v)$, we preserve the tokens the model is most certain about from $\hat{T}_0$, and revert the less certain ones back to the '[M]' token to form $T_{t_{k-1}}$.

This iterative refinement of prediction and remasking continues until the final timestep is reached. The resulting sequence at this point, now fully "denoised", is produced as the final textual description. This method allows GeoDiT to consider the global context of both the image and the entire text sequence at every step, leading to more coherent and globally consistent outputs.

3.4. Implementation Details

Our framework is implemented using the PyTorch library and trained on NVIDIA H200 GPUs. Below we detail the specific configurations and hyperparameters for our model's components, training, and inference stages.

Model Configurations. The key components of GeoDiT are configured as follows. The Visual Backbone is a SigLIP-2 (ViT-SO400M) model, which outputs visual embeddings with a hidden dimension of $D_v = 1152$. The MLP Projector consists of two linear layers with a GELU activation function in between, projecting the visual embeddings from D_v to the generative core's hidden dimension d. Our Generative Core, initialized from LLaDA-8B, is a 32-layer bidirectional Transformer with a hidden dimension of $d = 4096$ and 32 attention heads.

Training Details. The two-stage training process utilizes the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. For Stage I (Vision-Language Alignment), we train the MLP projector for 1 epoch on the SkyScript [36] dataset. The global batch size is 96. We employ a cosine learning rate scheduler, which includes a warm-up phase over the first 3% of the total training steps to a peak learning rate of 1×10^{-3}, after which the learning rate decays. No weight decay is applied in this stage. For Stage II (Full Instruction Tuning), the entire model is fine-tuned for 1 epochs on our curated MMRS-1M [46] optical subset. The global batch size for this stage is 24. We again utilize a cosine learning rate scheduler with a warm-up phase over the first 3% of training steps, but with a lower peak learning rate of 1×10^{-5}. No weight decay is used.

Inference Details. During inference, the iterative generation process is performed over $N = 8$ discrete timesteps. The maximum length L for the generated response is set to 16 tokens for image captioning, 32 tokens for object detection and 8 tokens for others. We employ the low-confidence remasking strategy for iterative refinement. All reported results are generated using greedy decoding and without classifier-free guidance.

4. Experiments

This section presents a comprehensive set of experiments designed to validate the central hypothesis of this work: that reframing geospatial language generation as a parallel, iterative refinement process offers a structurally superior alternative to the conventional autoregressive paradigm. To this end, we first detail the experimental setup, including the datasets, evaluation metrics, and baselines (Section 4.1).

Method	UCM-Captions			RSICD			NWPU-Captions			Sydney-Captions		
	B-4	MT	Cr	B-4	MT	Cr	B-4	MT	Cr	B-4	MT	Cr
Commercial Autoregressive Vision-Language Models												
GPT-4V [1]	7.4	18.7	34.8	5.4	14.8	30.9	15.2	22.9	45.6	9.1	18.5	24.0
Claude-4 (Anthropic 2025)	4.3	11.7	22.2	2.1	10.6	19.8	11.6	18.9	42.8	5.7	15.5	21.6
Open-source Diffusion-based Vision-Language Models												
LLaDA-V [42]	22.8	17.8	38.6	9.8	10.9	31.4	12.3	13.6	29.8	16.4	12.8	24.3
LaVida [19]	24.0	16.2	42.4	12.5	11.2	30.1	15.0	16.9	33.8	19.4	16.7	32.9
MMaDA [40]	25.6	18.9	43.6	15.8	13.9	31.2	14.6	15.7	31.1	20.2	19.6	32.8
Open-source Autoregressive Vision-Language Models												
LLaVA-1.5 [23]	24.6	26.7	47.7	21.3	16.4	43.6	47.6	26.5	56.9	39.1	25.3	22.6
Qwen2.5-VL [2]	23.1	30.5	35.9	13.9	20.6	35.3	24.1	26.4	47.8	16.1	22.5	27.9
GeoChat [17]	21.0	20.8	36.8	15.8	14.1	34.4	37.8	25.2	44.2	24.6	16.5	20.4
VHM [29]	42.4	26.9	64.2	19.5	22.2	53.5	47.9	18.9	40.4	37.0	21.5	33.2
EarthDial [32]	14.6	18.1	35.9	27.8	25.4	115.3	35.9	28.8	69.3	45.0	39.4	113.0
GeoDiT (ours)	**44.7**	**32.9**	**73.8**	**28.6**	**26.8**	**135.6**	**62.2**	**28.9**	**77.4**	**47.2**	**40.8**	**128.3**

Table 2. Quantitative comparison on remote sensing captioning benchmarks. Metrics are grouped to assess distinct qualities: narrative-based (BLEU4, METEOR) and object-centric (CIDEr). Best performance is in bold. Note GeoDiT's significant and consistent advantage on the object-centric CIDEr metric.

Method	DIOR-RSVG		VRSBench		AVVG		RSVG	
	VG	DET	VG	DET	VG	DET	VG	DET
Commercial Autoregressive Vision-Language Models								
GPT-4V	3.7	-	0	-	0	-	1.5	-
Claude-4	9.0	-	0.1	-	0	-	2.0	-
Open-source Diffusion-based Vision-Language Models								
LLaDA-V	0.0	-	0.0	-	0.0	-	0.0	-
LaVida	0.0	-	0.0	-	0.0	-	0.0	-
MMaDA	0.0	-	0.0	-	0.0	-	0.0	-
Open-source Autoregressive Vision-Language Models								
LLaVA-1.5	11.4	2.2	15.4	3.8	0.5	0.0	10.5	2.8
Qwen2.5-VL	36.3	17.9	45.2	19.6	0.5	0.0	1.0	0.0
GeoChat	31.4	1.4	56.3	3.4	7.5	0.4	5.5	0.0
VHM	55.9	3.4	33.9	2.8	0.0	0.0	2.5	0.0
EarthDial	46.1	13.1	14.4	0.7	3.5	0.0	42.0	13.5
GeoDiT (Ours)	**60.4**	**20.8**	**63.7**	**24.9**	**21.7**	**11.4**	**43.2**	**18.7**

Table 3. Quantitative comparison on visual grounding (VG) and object detection (DET) benchmarks.

Method	RSVQA-LR			RSVQA-HR		Classification	
	LR-R	LR-P	LR-C	HR-A	HR-C	AID	WHU-RS19
Commercial Autoregressive Vision-Language Models							
GPT-4V	74.3	81.3	56.7	18.4	33.6	77.3	94.7
Claude-4	63.4	79.6	65.4	21.7	42.8	45.5	66.1
Open-source Diffusion-based Vision-Language Models							
LLaDA-V	70.3	62.7	18.6	36.3	34.8	56.4	62.3
LaVida	66.8	58.1	14.2	26.6	44.3	55.8	60.8
MMaDA	71.6	60.8	16.8	28.9	37.6	57.6	59.7
Open-source Autoregressive Vision-Language Models							
LLaVA-1.5	75.5	57.1	70.6	0	66.0	47.3	62.1
Qwen2.5-VL	94.3	62.5	43.1	0	53.4	63.8	79.5
GeoChat	96.2	89.3	88.2	23.3	75.7	71.5	89.5
VHM	90.6	80.4	86.3	24.3	79.6	79.0	91.8
EarthDial	66.0	64.3	87.3	34.9	71.8	62.5	73.4
GeoDiT (ours)	**98.1**	**91.1**	**90.2**	**37.6**	**80.6**	**81.2**	**95.0**

R/P/C/A: Rural / Presence / Comparison / Area

Table 4. Performance evaluation on remote sensing visual question answering (RSVQA) and scene classification.

We then present the main quantitative results on a diverse suite of tasks, including image captioning , visual grounding , and visual question answering (Section 4.2). Following this, we conduct a series of ablation studies to dissect the model and analyze the contribution of its key components (Section 4.4). Finally, we provide a qualitative analysis with visual examples to offer further intuitive insights into the model's behavior and advantages (Section 4.3).

4.1. Experimental Setup

Datasets We evaluate our model on a wide array of public benchmarks to assess its performance across diverse geospatial tasks. For remote sensing captioning, we use four standard datasets: UCM-Captions [30], RSICD [26], NWPU-Captions [5], and Sydney-Captions [30]. For visual grounding and object detection, our evaluation is conducted on DIOR-RSVG [44], VRSBench [21], AVVG [50], and RSVG [34]. For visual question answering and scene classification, we employ the RSVQA-LR/HR [24], AID [38], and WHU-RS19 [37] benchmarks. This comprehensive suite of benchmarks ensures a thorough and robust evaluation of GeoDiT's capabilities and its generalization across different tasks and data domains.

Evaluation Metrics. We employ a set of standard metrics to evaluate performance across the different tasks. For captioning, we report scores for BLEU-4 (B-4), METEOR (MT), and CIDEr (Cr). We place particular emphasis on the CIDEr score, as it is specifically designed to measure consensus in object-centric descriptions, which directly aligns with our core hypothesis regarding the generation of structured, non-narrative semantics. With an IoU threshold of 0.5, we use Accuracy (Acc@0.5) for visual grounding (VG) task and mean Average Precision (mAP@0.5) for object detection (DET) task. For all visual question answering and classification tasks, we generate responses using greedy decoding and report accuracy ultilizing BERT-based similar-

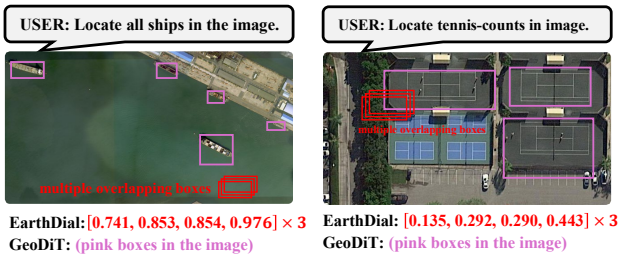


Figure 4. Qualitative examples of the generative failure mode in autoregressive models for object detection

ity.

Compared Methods. We conduct a comprehensive comparison of GeoDiT against a wide range of state-of-the-art vision-language models. To provide a clear overview of the competitive landscape, we group these baselines into three main categories: 1) **Commercial Autoregressive Models**: Large-scale, proprietary systems such as GPT-4V and Claude-4 [10] that represent the upper echelon of general VLM capabilities. 2) **Open-source Diffusion-based Models**: This category includes LLaDA-V, LaVida, and MMaDA. While they operate on similar non-autoregressive principles, they are designed for general-domain tasks and not specifically architected for geospatial semantics. 3) **Open-source Autoregressive Models**: This group comprises the dominant paradigm in academic research and includes leading remote sensing VLMs like LLaVA-1.5, Qwen2.5-VL, GeoChat, VHM, and EarthDial. This curated set of baselines allows for a multi-faceted evaluation of our model against both structurally similar and fundamentally different state-of-the-art approaches.

4.2. Main Results

We now present the main results of our experiments, evaluating GeoDiT's performance on a diverse set of tasks and comparing it against state-of-the-art models.

Image Captioning. The results for remote sensing image captioning are presented in Table 2. On the evaluated benchmarks, GeoDiT establishes a new state-of-the-art for the BLEU-4, METEOR, and CIDEr metrics. For instance, on the RSICD benchmark, GeoDiT achieves a CIDEr score of 135.6, marking a 17.6% relative improvement over the strongest competitor (EarthDial). Similarly, on Sydney-Captions, our model's score of 128.3 represents a 13.5% relative gain over the runner-up. This consistent and dramatic outperformance on CIDEr, a metric specifically designed to evaluate consensus on the set of contained objects, provides powerful validation for our thesis, highlighting a fundamental alignment between GeoDiT's parallel refinement process and the nature of unordered geospatial descriptions.

Visual Grounding and Detection. For tasks that demand precise spatial localization, namely visual grounding (VG) and object detection (DET), GeoDiT demonstrates a comprehensive superiority. As shown in Table 3, GeoDiT sets a new state-of-the-art across all benchmarks and on both evaluation metrics by a significant margin. For instance, on the challenging VRSBench dataset, GeoDiT achieves a VG score of 63.7 and a DET score of 24.9, substantially outperforming the strongest baselines in each category. This gap stems from a fundamental limitation inherent to sequential generation. Autoregressive models, constrained by their token-by-token process, exhibit a critical failure mode when tasked with enumerating multiple objects. The generation process of each subsequent bounding box is conditioned on the preceding ones, creating a path dependency inducing the premature convergence on a single salient object, manifested as the repeated generation of duplicate or minimally perturbed coordinates, precluding the identification of other distinct entities (see Fig. 4). Consequently, their ability to perform a comprehensive object canvass is structurally impaired, leading to a severe degradation in DET performance.

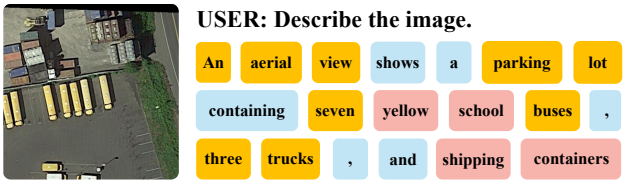


Figure 5. Visualization of the hierarchical generation process of GeoDiT. The color of each token corresponds to its relative finalization step during the iterative inference process: yellow indicates early-stage tokens, pink indicates middle-stage tokens, and blue indicates late-stage tokens.

Visual Question Answering and Classification. As presented in Table 4, GeoDiT's capabilities are demonstrated across both visual question answering and scene classification, where it establishes a new state-of-the-art on all evaluated sub-tasks and datasets. The model's proficiency is consistent across the full suite of VQA tasks, which require parsing complex spatial relationships and object attributes. For instance, GeoDiT scores 98.1 on the RSVQA-LR-R (Rural) sub-task and 80.6 on the demanding RSVQA-HR-C (Comparison) task. This strong performance extends to scene classification, where GeoDiT also surpasses prior methods, reaching an accuracy of 95.0 on WHU-RS19. This consistent success across a spectrum of tasks—from answering compositional queries about spatial arrangements (VQA) to holistic scene recognition (Classification)—offers a key insight into the model's underlying mechanism. This suggests that the parallel refinement paradigm provides a more fundamental advantage than being merely a specialized tool for structured outputs. By processing the scene globally and iteratively, the model excels not only at disentangling complex, localized relationships but also at forming the cohesive, scene-level understanding required for single-label classification.

Masking Strategy	BLEU-4 (RSICD)	CIDEr (RSICD)	mAP@0.5 (DIOR-RSVG)	Acc. (AID)
Random Remasking	27.3	121.8	15.5	63.4
Low-Confidence (Ours)	**28.6**	**135.6**	**20.8**	**67.6**
Improvement (%)	↑ 4.76%	↑ 11.3%	↑ 34.2%	↑ 6.21%

Table 5. Ablation study on the masking strategy. We compare our proposed low-confidence remasking against a random remasking baseline.

#timesteps (N)	BLEU-4 (RSICD)	CIDEr (RSICD)	mAP@0.5 (DIOR-RSVG)	Acc. (AID)
1	21.0	65.8	7.5	76.5
2	25.3	105.1	14.2	79.8
4	27.8	127.3	18.9	70.7
8	28.6	135.6	20.8	81.2
16	28.7	136.2	21.1	81.3

Table 6. Ablation on inference timesteps (N). Performance saturates at $N = 128$, offering the optimal trade-off between generation quality and computational cost for structured outputs.

4.3. Qualitative Analysis

To provide an intuitive understanding of the quantitative results presented in Section 4.2, we now visualize and analyze several qualitative examples. These cases are selected to highlight the practical differences in behavior between our parallel refinement paradigm and conventional autoregressive models on key geospatial tasks.

Analysis of Structural Robustness. Figure 4 provides a compelling visual demonstration of a critical failure mode inherent in autoregressive models during detection tasks. In both the harbor and tennis court scenes, the baseline model generates an initial, misplaced bounding box—for instance, locating a section of the pier instead of a ship. Critically, the model then becomes trapped in a generative loop, fixating on its initial error and repeatedly outputting the same incorrect coordinates. This behavior, visualized by the overlapping red boxes on an irrelevant location, leads to a catastrophic failure on mAP. This qualitative evidence strongly supports our central claim: the parallel refinement paradigm of GeoDiT is structurally more robust for multi-object understanding, as it avoids the sequential dependencies that lead to such error-compounding generative failures.

Hierarchical Generation Process. Figure 5 visualizes the non-autoregressive generation process of GeoDiT. It reveals a distinct, hierarchical generation pattern that fundamentally differs from linear autoregressive models. The model first establishes the macroscopic scene context (e.g., 'An aerial view shows a parking lot') and identifies the primary semantic anchors such as key objects and their counts ('seven', 'buses', 'three', 'trucks') in the early stages (yellow). Subsequently, it refines the description by adding attributes ('yellow', 'school', 'shipping containers') in the middle stages (pink). Finally, it completes the sentence by inserting the necessary grammatical and syntactical components ('containing', 'and', '.') in the late stages (blue). This 'context-first, entity-second, syntax-last' process is only possible through a parallel, holistic understanding of the image and strongly validates our core paradigm.

4.4. Ablation Studies

Analysis of Masking Strategy. Table 5 presents the ablation study on our masking strategy, confirming that our low-confidence approach is decisively superior to a random remasking baseline. This superiority, however, is not uniform and is most pronounced on tasks requiring high precision. We observe a remarkable 34.2% improvement in object detection (mAP) and a significant 11.3% gain on the object-centric CIDEr metric. In contrast, the gains are more moderate for BLEU-4 (↑ 4.76%) and simple VQA accuracy (↑ 6.21%). This disparity is informative and expected. By focusing on uncertain tokens, our intelligent masking strategy provides the greatest benefit when refining high-stakes details like precise bounding box coordinates and key object nouns—the very elements that are critical for high mAP and CIDEr scores. This confirms our design choice is a key contributor to generating high-fidelity structured outputs.

Analysis of Inference Timesteps. To investigate the contribution of our iterative refinement process, we ablate the number of inference timesteps, N, with results detailed in Table 6. A highly informative pattern emerges: performance on metrics sensitive to structural and object-centric correctness, namely CIDEr and mAP, scales steeply with the initial steps ($N = 1, \cdots, 16$), underscoring the necessity of a multi-step approach to resolve the parallel semantics of geospatial scenes where single-pass generation inherently falters. This trend is sharply contrasted by the far less sensitive scene classification task (Acc.), whose performance plateaus much earlier, confirming that the primary value of our method lies in meticulously constructing structured outputs rather than making a single holistic judgment. As performance across all tasks begins to saturate at $N = 8$, yielding only marginal gains when doubling the computational budget to $N = 16$, we adopt $N = 8$ as our default setting, a robust and efficient operating point that captures the vast majority of the model's capabilities without incurring unnecessary inference latency.

5. Conclusion

This work addresses a fundamental limitation of the prevailing autoregressive paradigm: its structural misalignment with the parallel nature of remote sensing data. To overcome this, we introduced GeoDiT, a novel diffusion-based vision-language model that operationalizes a parallel, iterative refinement process. Extensive experiments show GeoDiT establishes a new state-of-the-art on object-centric and structured-output tasks, such as multi-object detection and image captioning, validating the superior-

ity of our structurally-aligned approach. Our work underscores the critical importance of aligning the generative architecture with the data's inherent structure and presents a promising new direction for complex geospatial intelligence.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 6

[3] TC Chakraborty, Zander S Venter, Matthias Demuzere, Wenfeng Zhan, Jing Gao, Lei Zhao, and Yun Qian. Large disagreements in estimates of urban land across scales and their implications. *Nature communications*, 15(1):9165, 2024. 1

[4] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 4

[5] Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. Nwpu-captions dataset and mlca-net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 4, 6

[6] Dengxin Dai and Wen Yang. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geoscience and remote sensing letters*, 8(1):173–176, 2010. 4

[7] Yanghua Di, Zhiguo Jiang, Haopeng Zhang, and Gang Meng. A public dataset for ship classification in remote sensing images. In *Image and Signal Processing for Remote Sensing XXV*, pages 515–521. SPIE, 2019. 4

[8] Yanghua Di, Zhiguo Jiang, and Haopeng Zhang. A public dataset for fine-grained ship classification in optical remote sensing images. *Remote Sensing*, 13(4):747, 2021. 4

[9] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7778–7796, 2021. 4

[10] Maxim Enis and Mark Hopkins. From llm to nmt: Advancing low-resource machine translation with claude, 2024. 7

[11] Jing Gao and Brian C O'Neill. Mapping global urban land for the 21st century with data-driven simulations and shared socioeconomic pathways. *Nature communications*, 11(1): 2302, 2020. 1

[12] Junwei Han, Dingwen Zhang, Gong Cheng, Lei Guo, and Jinchang Ren. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3325–3337, 2014. 4

[13] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, pages 37–55. Springer, 2024. 3

[14] 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. 4

[15] Siwen Jiao, Yangyi Fang, Baoyun Peng, Wangqun Chen, and Bharadwaj Veeravalli. Lavida drive: Vision-text interaction vlm for autonomous driving with token selection, recovery and enhancement, 2025. 2

[16] Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, et al. Mercury: Ultra-fast language models based on diffusion. *arXiv e-prints*, pages arXiv–2506, 2025. 2

[17] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024. 1, 2, 6

[18] Ke Li, Gong Cheng, Shuhui Bu, and Xiong You. Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2337–2348, 2017. 4

[19] Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. Lavida: A large diffusion language model for multimodal understanding, 2025. 2, 6

[20] Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124:103497, 2023. 1

[21] Xiang Li, Jian Ding, and Mohamed Elhoseiny. Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding. In *Advances in Neural Information Processing Systems*, pages 3229–3242. Curran Associates, Inc., 2024. 6

[22] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 6

[24] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (12):8555–8566, 2020. 4, 6

[25] Yang Long, Yiping Gong, Zhifeng Xiao, and Qing Liu. Accurate object localization in remote sensing images based on

convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2486–2498, 2017. 4

[26] 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. 4, 6

[27] Lidong Mo, Constantin M Zohner, Peter B Reich, Jingjing Liang, Sergio De Miguel, Gert-Jan Nabuurs, Susanne S Renner, Johan Van Den Hoogen, Arnan Araza, Martin Herold, et al. Integrated global assessment of the natural forest carbon potential. *Nature*, 624(7990):92–101, 2023. 1

[28] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. 4, 5

[29] Chao Pang, Xingxing Weng, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Shuai Wang, Litong Feng, Gui-Song Xia, et al. Vhm: Versatile and honest vision language model for remote sensing image analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6381–6388, 2025. 1, 2, 6

[30] Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. Deep semantic understanding of high resolution remote sensing image. In *2016 International conference on computer, information and telecommunication systems (Cits)*, pages 1–5. IEEE, 2016. 4, 6

[31] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. 4

[32] Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, et al. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14303–14313, 2025. 2, 6

[33] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 4

[34] Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International conference on Multimedia*, pages 404–412, 2022. 6

[35] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 3

[36] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5805–5813, 2024. 5

[37] Gui-Song Xia, Wen Yang, Julie Delon, Yann Gousseau, Hong Sun, and Henri Maître. Structural high-resolution satellite image indexing. In *ISPRS TC VII Symposium-100 Years ISPRS*, pages 298–303, 2010. 6

[38] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 6

[39] Jun Yang, Peng Gong, Rong Fu, Minghua Zhang, Jingming Chen, Shunlin Liang, Bing Xu, Jiancheng Shi, and Robert Dickinson. The role of satellite remote sensing in climate change studies. *Nature climate change*, 3(10):875–883, 2013. 1

[40] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models, 2025. 2, 6

[41] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 4

[42] Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning, 2025. 2, 6

[43] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *arXiv preprint arXiv:2204.09868*, 2022. 4

[44] Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13, 2023. 4, 6

[45] Meimei Zhang, Fang Chen, and Bin Li. Multistep question-driven visual question answering for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023. 4

[46] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. 5

[47] Yuanlin Zhang, Yuan Yuan, Yachuang Feng, and Xiaoqiang Lu. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57 (8):5535–5548, 2019. 4

[48] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[49] Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. Mutual attention inception network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 4

[50] Yue Zhou, Mengcheng Lan, Xiang Li, Yiping Ke, Xue Jiang, Litong Feng, and Wayne Zhang. Geoground: A unified large vision-language model for remote sensing visual grounding, 2025. 6

[51] Haigang Zhu, Xiaogang Chen, Weiqun Dai, Kun Fu, Qixiang Ye, and Jianbin Jiao. Orientation robust object detection in aerial images using deep convolutional neural network. In *2015 IEEE international conference on image processing (ICIP)*, pages 3735–3739. IEEE, 2015. 4

[52] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and remote sensing letters*, 12(11):2321–2325, 2015. 4

GeoDiT: A Diffusion-based Vision-Language Model for Geospatial Understanding

Supplementary Material

6. Implementation Details of Remasking Strategy

The low-confidence remasking strategy is the core mechanism for iterative refinement in the non-autoregressive inference process of our model, GeoDiT. This strategy is designed to dynamically focus the model's generative capacity on the most uncertain parts of a prediction while preserving high-confidence tokens that have already been determined. This progressive refinement facilitates the generation of globally consistent and structured outputs, following effective practices established in discrete diffusion models.

The inference process commences with a template T_N composed entirely of `[M]` tokens and proceeds over N discrete timesteps. At each iteration k (from timestep t_k to t_{k-1}), the strategy is executed as follows:

1. **Full Sequence Prediction:** The model, p_θ, takes the masked sequence at the current step, T_{t_k}, and the visual condition vectors, C_v, as input. It produces a probability distribution over the entire vocabulary for each position in the sequence, $p_\theta(T_0|T_{t_k}, C_v)$. From this distribution, a provisional but complete sequence prediction, $\hat{T}_0$, is generated by selecting the token with the maximum probability (i.e., argmax) at each position.
2. **Confidence Score Acquisition:** For each predicted token $\hat{T}_0^i$ at position i in the provisional sequence $\hat{T}_0$, its confidence score is defined as the probability assigned to it by the model, $p_\theta(T_0^i = \hat{T}_0^i|T_{t_k}, C_v)$. This score directly reflects the model's certainty for the prediction at that specific position.
3. **Determining the Number of Tokens to Remask:** The number of tokens to be remasked at step k, denoted m_k, is determined by a predefined scheduling function $\gamma(t)$. This function maps the current timestep t_k (which anneals from 1 towards 0) to a ratio that dictates the percentage of the sequence to be masked. Specifically, $m_k = \lceil \gamma(t_k) \cdot L \rceil$, where L is the total sequence length. We employ a cosine schedule for $\gamma(t)$, which results in a higher masking ratio during the early stages of inference (when t_k is large) and a lower ratio in the later stages (when t_k is small). This facilitates a coarse-to-fine refinement process.
4. **Selecting and Applying the Mask:** Based on the confidence scores calculated in step 2, the model identifies the m_k tokens in $\hat{T}_0$ with the lowest scores. These are deemed the "low-confidence" tokens. The positions of these tokens are then reverted to the special `[M]` token, while all other high-confidence tokens from $\hat{T}_0$ are preserved. The resulting sequence, $T_{t_{k-1}}$, serves as the input for the subsequent iteration $k-1$.

This cyclic process of prediction and remasking allows the model to make judgments based on global information at every step, effectively avoiding the cascading errors that can occur in sequential, autoregressive generation. Our ablation study (see Table 4 in the main paper) confirms that this intelligent masking strategy is crucial for enhancing performance on tasks that demand high-precision details, such as bounding box coordinates and key object nouns, making it a key factor in achieving high-fidelity structured outputs.

7. Training Configurations

7.1. Stage I: Vision-Language Alignment

The initial vision-language alignment was conducted by training only the MLP projector, keeping the vision and language model backbones frozen[cite: 188]. The specific hyperparameters used, based on the provided training script, are detailed below.

- **Optimizer:** AdamW, with $\beta_1 = 0.9$ and $\beta_2 = 0.95$.
- **Learning Rate:** A peak learning rate of 1×10^{-3} was used.
- **Learning Rate Scheduler:** We employed a cosine decay schedule, with a warm-up phase over the first 3% of the total training steps.
- **Weight Decay:** 0.0
- **Training Epochs:** The alignment was performed for a single epoch.
- **Global Batch Size:**96. This was achieved with a per-device batch size of 16 and 1 gradient accumulation step, distributed across 6 GPUs.
- **Precision:** Training utilized BF16 and was accelerated with TF32 enabled.
- **Model Configuration:** The maximum model length was set to 8192 tokens. Gradient checkpointing was enabled to conserve memory. We used the 'sdpa' attention implementation.
- **Infrastructure:** The training was managed using DeepSpeed with a ZeRO Stage 3 configuration.

7.2. Stage II: Full Instruction Tuning

Following the initial alignment, the model underwent end-to-end fine-tuning on a large-scale, instruction-formatted remote sensing dataset. All major components of the model

were unfrozen for this stage. The specific hyperparameters are detailed below.

- **Training Objective:** All primary model components (vision tower, MLP projector, and language model) were unfrozen and trained end-to-end. The model was initialized with the MLP projector weights obtained from Stage I.
- **Learning Rates:** A differential learning rate scheme was used. The main model components were trained with a peak learning rate of 1×10^{-5}, while the vision tower was fine-tuned with a lower learning rate of 2×10^{-6}.
- **LR Scheduler & Weight Decay:** Consistent with Stage I, a cosine decay scheduler with a 3% warmup ratio and a weight decay of 0.0 were used.
- **Global Batch Size:** 24. This was configured with a per-device batch size of 2, distributed across 6 GPUs with 2 gradient accumulation steps.
- **Training Epochs:** The model was fine-tuned for a single epoch on the instruction dataset.
- **Image Handling:** To process high-resolution imagery efficiently, we utilized an 'anyres' strategy. This involved variable grid pinpoints (from 1×1 up to 6×6) to create a maximum of 4 image patches per sample.
- **Infrastructure & Optimization:** The training setup leveraged DeepSpeed ZeRO Stage 3 and PyTorch's 'sdpa' attention implementation. For further optimization, the model was compiled using 'torch.compile' with the 'inductor' backend.
- **Precision & Model Configuration:** BF16 precision, TF32 acceleration, a maximum model length of 8192, and gradient checkpointing were enabled, consistent with the settings in Stage I.

8. Visualizations for the generation process of GeoDiT

To further substantiate the unique, non-autoregressive generation paradigm of GeoDiT, this section provides additional visualizations that complement the analysis presented in Figure 5 of the main paper. The examples below, featuring diverse geospatial scenes, demonstrate that the hierarchical, coarse-to-fine generation process is a consistent and fundamental characteristic of our model's behavior, rather than an isolated phenomenon.

Across all presented cases, GeoDiT consistently exhibits a multi-stage, parallel refinement strategy. As illustrated in Figures 6, 7, and 8, the model first establishes the macroscopic scene context and key semantic anchors. These foundational elements, such as the primary location type (e.g., 'parking lot', 'airport') and the principal objects and their counts (e.g., 'seven', 'buses', 'airplanes'), are typically finalized in the early stages of the diffusion process (indicated by yellow tokens).

Subsequently, the model dedicates the middle stages

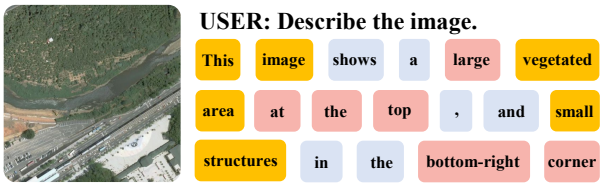


Figure 6. Visualization of the Hierarchical Generation Process for a Scene with Mixed Land Use.

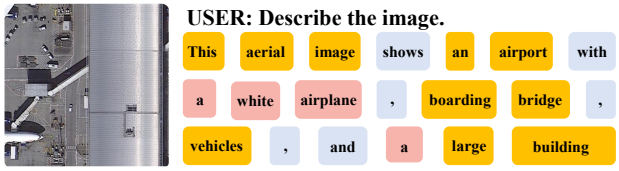


Figure 7. Visualization of the Hierarchical Generation Process for an Airport Scene.

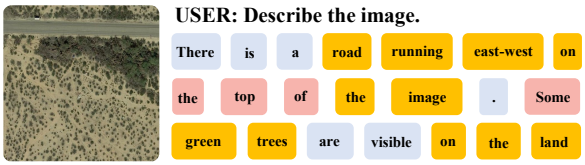


Figure 8. Visualization of the Hierarchical Generation Process for a Rural Landscape.

to refining the description by adding attributes and secondary entities (pink tokens), such as descriptive adjectives ('yellow', 'large') or related objects ('shipping containers'). Finally, in the late stages, the model inserts the necessary grammatical and syntactical components ('containing', 'and', '.') to form a coherent and complete sentence (blue tokens). This coarse-to-fine process, which relies on a holistic understanding of the entire image-text relationship at every step, is structurally impossible for linear, token-by-token autoregressive frameworks.