Fine-Grained Remote Sensing Imagery Generation Driven by Expert Knowledge and Hierarchical Captions

Jiaxin Ren[1,2,3], Wanzeng Liu[4,*], Feng Zhang[3], Jun Chen[1,4,*], Jiadong Zhang[5], Shunxi Yin[6], Shaoxuan Zhao[5], Guanfan Xi[1], Di Chen[1], Kuanlin Dong[1]

[1]Moganshan Geospatial Information Laboratory, Huzhou, China
[2]Key Laboratory of Monitoring, Evaluation and Early Warning of Territorial Spatial Planning Implementation, Ministry of Natural Resources, Chongqing, China
[3]School of Earth Sciences, Zhejiang University, Hangzhou, China
[4]National Geomatics Center of China, Beijing, China
[5]School of Geosciences and Info-Physics, Central South University, Changsha, China
[6]School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou, China
(jaycecd@foxmail.com, lwz@ngcc.cn, zfcarnation@zju.edu.cn, chenjun@ngcc.cn, zjdgis@csu.edu.cn, yshunxi@gmail.com, zsx2020@csu.edu.cn, guanfanx001@gmail.com, geodichen@gmail.com, dkuanlin@163.com)

Keywords: Remote Sensing Scene Generation, Hierarchical Captioning, Expert Knowledge, Text-to-Image, Diffusion Model.

Abstract

Current diffusion models struggle to achieve fine-grained remote sensing imagery (RSI) generation. This limitation fundamentally stems from their reliance on "flattened" text prompts, which overlook the inherent hierarchical structure of RSI. This paper proposes a fine-grained RSI generation method driven by expert knowledge and hierarchical captions. We first deconstruct RSI into a hierarchical "element-relation-scene" caption and employ an automatic caption optimization mechanism, grounded in spatial relation knowledge, to ensure high fidelity. Critically, we introduce a novel hierarchical caption encoding mechanism that efficiently injects decoupled hierarchical caption segments into the U-Net's cross-attention layers. This design enables the model to exert hierarchical and decoupled attentional control over the global scene, spatial layout, and geographical element details during the denoising process. Experiments demonstrate that, when combined with efficient fine-tuning algorithms such as LoRA, our method significantly outperforms traditional single-level captions across all six evaluation metrics, exemplified by the FID metric decreasing from 228.43 to 205.59 and the GSHPS metric increasing from 0.86 to 0.92. This research provides a new paradigm for controllable remote sensing scene generation, establishing an effective link between hierarchical semantic understanding and the progressive generation process of diffusion models.

1. Introduction

Text-to-Image (T2I) diffusion models have achieved revolutionary progress in generating high-fidelity and diverse natural images. This success is now driving their application in specialized domains such as remote sensing. However, while T2I models excel at simulating natural image features, they face severe challenges in the fine-grained generation of remote sensing imagery (RSI), a task demanding precise spatial layouts and specific combinations of geographical elements. Existing T2I models in remote sensing applications often struggle to accurately control the position, quantity, or topological relationships of key geographical elements, leading to results that are geospatially unreasonable or semantically distorted. This bottleneck limits the practical application potential of RSI generation. Whether serving as a data augmentation strategy to provide high-quality synthetic samples for downstream tasks such as object detection and semantic segmentation, or enabling rapid generation of spatially controllable scene previews from textual plans in urban planning, there is an urgent demand for fine-grained control capabilities in generative models.

We argue that this bottleneck stems from the fundamental reliance of T2I models on "flattened" or "single-level" text prompts. Standard natural language prompts (e.g., "*a bird's-eye view of a factory*") are semantically ambiguous; they lack an explicit decomposition of the complex scene's internal structure. A remote sensing scene (RSS) is inherently hierarchical, composed of a global scene concept (e.g., "*industrial area*"), a set of specific geographical elements (e.g., "*factory*", "*chimney*", "*road*"), and the precise spatial relations among them (e.g., "*road surrounds factory*"). Flattened prompts cannot provide this structured, composable guidance for the generation process, thereby limiting the precision of the model's output.

To overcome this limitation, this paper proposes an expert knowledge and hierarchical caption-driven method for fine-grained RSI generation. The method first deconstructs RSI into a hierarchical "element-relation-scene" triplet to provide structured guidance. We utilize a hybrid model system to automatically extract information across these three dimensions (Ren et al., 2021; Liu et al., 2022; Liu et al., 2024; Ren et al., 2025; Zhang et al., 2026). We also introduce expert knowledge to correct noise induced by cross-domain models, thereby ensuring the high fidelity of the hierarchical captions. Critically, we propose a novel hierarchical caption encoding mechanism to efficiently inject this structured hierarchical caption into the diffusion model. This enables the U-Net's cross-attention mechanism to exert hierarchical and decoupled attentional control over the global scene, spatial layout, and geographical element details at each step of the denoising process. The model first establishes a global framework based on the "scene" concept. It subsequently constructs the key spatial structure using "relation" information and finally populates fine-grained geographical element details via the "element" caption. In this way, the model can explicitly learn the complex correspondence

* Corresponding Author

between the hierarchical caption and the generated imagery, thus achieving fine-grained control over the RSS.

Through experiments on multiple efficient fine-tuning algorithms, our model demonstrates its capacity to fully understand and utilize this structured, hierarchical input. Compared with methods using traditional single-level captions, our method achieves significant advantages on all key evaluation metrics, with generated RSS showing substantial improvements in fidelity, diversity, and semantic consistency. This research provides a new paradigm for controllable RSS generation, establishing an effective link between hierarchical semantic understanding and the progressive generation process of diffusion models.

2. Related Work

2.1 Natural Image Captioning

Natural Image Captioning (NIC) aims to generate grammatically correct and semantically accurate natural language descriptions for a given image. Early research in this domain was based on template matching and retrieval methods, subsequently shifting to encoder-decoder architectures. These architectures utilize convolutional neural networks (CNNs) to extract visual features and employ recurrent neural networks (RNNs) or their variants (e.g., LSTM) as decoders to generate sequential text. The introduction of attention mechanisms significantly improved descriptive accuracy by allowing the model to dynamically focus on different image regions while generating each word. Recent research paradigms have migrated towards large-scale vision-language pre-trained models (VLMs), such as those using CLIP's contrastive learning paradigm. By pre-training on massive image-text pairs, these models have achieved stronger zero-shot and few-shot generation capabilities, trending towards more controllable and fine-grained descriptions. However, these methods primarily generate single, holistic sentences. They lack the capability for explicit modeling of hierarchical structures or multi-scale relationships within complex scenes, failing to meet the structured knowledge input required for fine-grained scene generation.

2.2 Remote Sensing Image Captioning

The Remote Sensing Image Captioning (RSIC) task inherits the basic framework of natural image captioning but faces challenges unique to remote sensing images, such as vast differences in object scale, complex backgrounds, and a large volume of specialized geographical element terminology. Early RSIC models directly applied CNN-RNN architectures, which had been validated on natural images, to remote sensing data (Lu et al., 2018). To address the co-existence of multiple objects and complex spatial relations in remote sensing images, subsequent research focused on exploring multi-scale feature fusion, graph neural networks (GNNs), and more refined visual attention mechanisms. The recent trend is to build remote sensing-specific vision-language pre-trained models (RS-VLMs) or design specific modules to capture the fine-grained semantics and spatial relations of remote sensing images, thereby generating captions that contain richer geographical element details and precise spatial layouts. Although these methods have achieved progress on RSIC metrics, their generated captions remain semantically flattened.

2.3 Natural Image Generation

Image generation techniques aim to synthesize photorealistic and diverse images. Variational autoencoders (VAEs) and generative adversarial networks (GANs) are two major milestones in this field. VAEs excel at learning the latent distribution of data, but the generated images are prone to blurriness. GANs, through an adversarial process between a generator and a discriminator, achieved breakthroughs in the clarity and realism of generated images, with style-based generators (StyleGAN) performing exceptionally well in high-resolution face synthesis (Karras et al., 2019). Currently, diffusion models have become the dominant paradigm. By simulating a progressive denoising reverse process, diffusion models have surpassed GANs in both generation quality and diversity. For instance, denoising diffusion probabilistic models (DDPM) (Ho et al., 2020) achieve extremely high generation fidelity through a progressive denoising process. T2I models, such as the latent diffusion model (Rombach et al., 2022), have demonstrated the powerful conditional generation capabilities of diffusion models and have gradually formed the core of large-scale T2I models like DALL-E 2 (Ramesh et al., 2022) and Stable Diffusion 3.

2.4 Remote Sensing Image Generation

The core challenge in remote sensing image generation (RSIG) lies in simultaneously ensuring the spectral authenticity and geospatial reasonableness of the generated images. Initial efforts primarily relied on GANs and their conditional variants (Mirza and Osindero, 2014), which were used for synthesizing specific geographical elements (e.g., airplanes, ships) or for image style transfer. As the demand for scene controllability increased, research shifted towards using semantic maps or scene graphs as conditions to guide the generation process, aiming to control the categories and spatial layout of geographical elements. Inspired by the success of T2I models, recent RSIG research has begun to focus on text-driven remote sensing scene generation. This involves using diffusion models or GANs to take natural language captions as input to generate complex remote sensing scenes that conform to the text semantics (Ren et al., 2024, Tang et al., 2024). However, achieving precise control over a scene's specific geographical element attributes and fine-grained spatial relations remains a critical, unsolved problem.

3. Fine-Grained Remote Sensing Scene Generation Method Driven by Expert Knowledge and Hierarchical Captions

3.1 Hierarchical Caption Generation Fusing Expert Knowledge

3.1.1 Knowledge-Guided Remote Sensing Image Preprocessing: Condensed prior knowledge indicates that appropriate resolution and image size for deep learning model input are critical for constructing high-quality remote sensing scene datasets (Ren et al., 2024, Yin et al., 2025, Ren et al., 2026). Scaling remote sensing images often introduces discrepancies in recognition results. Therefore, this paper first performs channel selection and format conversion on the original multi-channel remote sensing images, converting them into three-channel RGB images. Preprocessing methods such as slicing, cropping, and padding are then applied to maintain the consistency of the preprocessed images. This ensures compatibility between the remote sensing images and deep learning algorithms, enhances model performance and generalization ability, and enables the model to accurately understand and analyze complex remote sensing scenes. Specifically, different preprocessing methods are selected based on the relationship between the original image size and input size, including large-size image slicing, medium-size image

cropping, and small-size image padding:

$$preprocessing=\begin{cases} slicing & \text{if } S_{image} \geq 2\times S_{input} \\ cropping & \text{if } S_{input} < S_{image} < 2\times S_{input} \\ padding & \text{if } S_{image} \leq S_{input} \end{cases} \quad (1)$$

Here, S_{image} represents the size of the remote sensing image I, denoted as $m\times n$; S_{input} represents the size of the remote sensing image data D input to the deep learning model, denoted as $s\times s$.

Through these preprocessing methods, the key information of the training samples becomes more distinct and conforms to the input specifications of deep learning models. This ensures the size and content of the images meet the model's training and inference requirements, effectively bridging the size disparity between remote sensing images and traditional deep learning input images. These methods provide suitable training data, ensure the applicability, consistency, and quality of the input data, and allow for better adaptation to complex geographic information scenarios, thereby improving model performance and generalization.

3.1.2 Automatic Generation of Hierarchical Captions for Remote Sensing Scenes Based on Spatial Relation Knowledge: To achieve automatic captioning of remote sensing scenes, Figure 1 illustrates the specific workflow of the proposed automatic generation method for hierarchical captions, which is based on spatial relation knowledge. This method employs multiple automatic captioning algorithms at different granularities for the same remote sensing image, including the geographical element-level Deep Danbooru (Liu et al., 2023) algorithm, the spatial relation-level BLIP-2 (Li et al., 2023) algorithm, and the scene concept-level EfficientNet (Tan and Le, 2019) algorithm. The resulting geographical element, spatial relation, and scene concept are organized according to an element-relation-scene structure. Spatial relation knowledge is then utilized to optimize the generated captions, ultimately achieving a precise, hierarchical cognition of the entire scene.

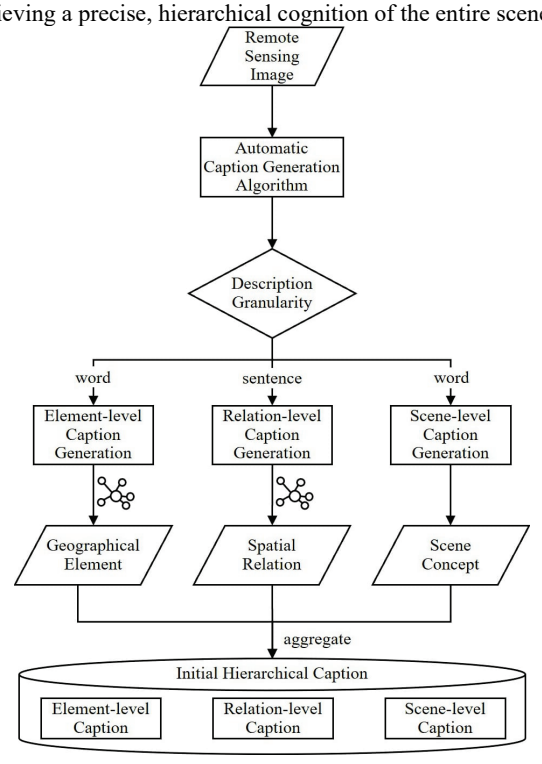

Figure 1. Automatic generation of hierarchical captions for remote sensing scenes based on spatial relation knowledge.

1) Automatic Extraction of Remote Sensing Scene Geographical Element Labels and Scene Concepts Based on Convolutional Neural Networks

For the automatic extraction of geographical element labels and scene concepts in remote sensing scenes, the complexity does not involve intricate spatial relation cognition. Therefore, directly utilizing convolutional neural networks for multi-label classification or semantic segmentation is sufficient to acquire the specific geographical element labels and scene concepts. This study employs the lightweight and efficient Deep Danbooru to extract geographical element labels from remote sensing scenes and utilizes EfficientNet for scene classification on remote sensing images to generate scene concepts.

Deep Danbooru identifies geographical elements by performing multi-label classification on the image, generating a series of words (e.g., color, shape, texture) as its initial caption. However, these labels lack specific subordinate or spatial relations and cannot reflect the overall functional effect of the geographic scene. This study utilizes EfficientNet fine-tuned on remote sensing scene images to achieve automatic cognition and scene concept extraction. Furthermore, if an image originates from an open-source dataset with existing class labels, that label can be directly adopted as the scene concept.

2) Automatic Generation of Remote Sensing Scene Spatial Relations Based on Spatial Relation Knowledge

For automatic spatial relation captioning, this study utilizes the BLIP-2 algorithm to generate a concise natural language caption for the image. This caption represents key geographical targets (e.g., mountains, rivers, cities, farmland) and their spatial relations (e.g., adjacent, across, contains).

A "golden rule" is to select captions that contain spatial relations. It is important to note that, unlike the side-view perspective of natural images, the vertical, nadir-looking perspective of remote sensing images almost never uses vertical spatial relations (e.g., underneath). Removing this class of spatial relations, which are common in natural image captions but nearly impossible in remote sensing images, effectively reduces descriptive inaccuracy and highlights the importance of prior knowledge (Zhang et al., 2025) . Therefore, by condensing expert knowledge, we define a spatial relation knowledge collection (SRKC). BLIP-2, using different decoding algorithms, is then employed to randomly generate multiple candidate text captions until a caption containing a spatial relation from the SRKC appears, at which point that caption is automatically selected.

In summary, the proposed automatic generation method for hierarchical captions, which is based on spatial relation knowledge, allows for the comprehensive acquisition of scene cognition results at different granularities. This process successfully transforms complex remote sensing images into structured, interpretable, multi-granularity text captions, providing high-quality input for subsequent fine-grained remote sensing image generation tasks.

3.2 Fine-Grained Remote Sensing Scene Generation Method Based on Hierarchical Captions and a Latent Space Diffusion Model

This paper proposes a fine-grained remote sensing scene generation method based on hierarchical captions and a latent space diffusion model. This method utilizes the generated

hierarchical caption dataset to efficiently fine-tune a diffusion model pre-trained on natural images. This approach achieves efficient remote sensing scene generation while remaining compatible with the visual model's original text-understanding capabilities. The overall workflow is illustrated in Figure 2.

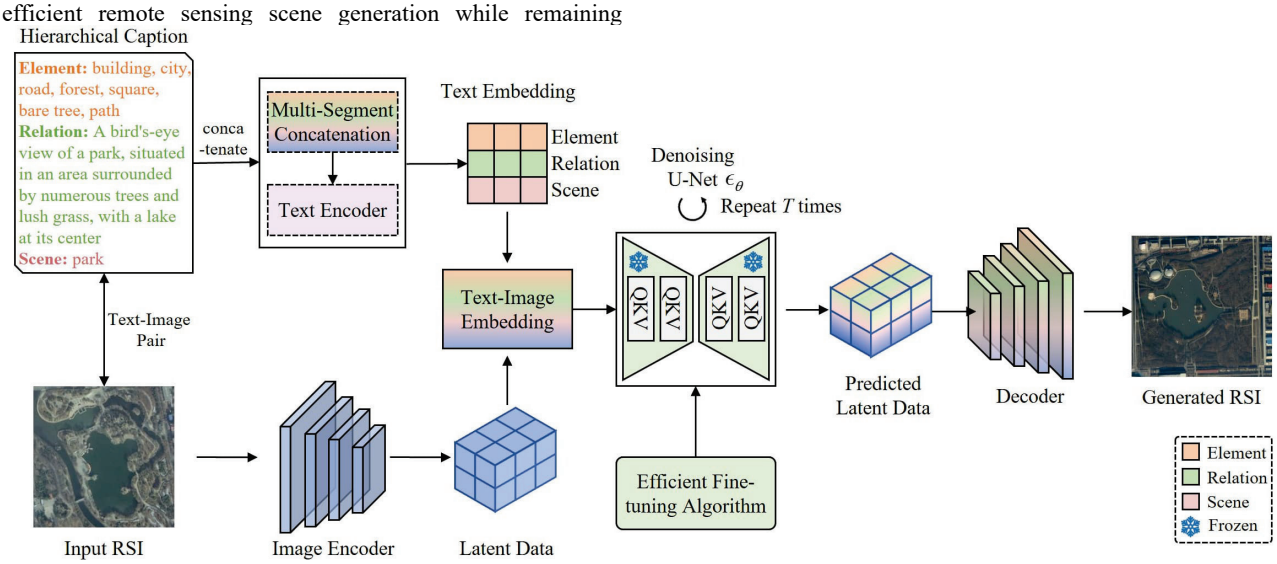

Figure 2 Fine-grained remote sensing scene generation method based on hierarchical captions and a latent space diffusion model.

3.2.1 Remote Sensing Scene Generation Model Based on Diffusion Models: Diffusion models are a class of generative models that generate data by simulating a progressive noising process and a denoising process. These models, initially derived from the thermodynamic diffusion process in physics, have achieved significant success in computer vision, particularly in natural image generation. The fundamental principle is: starting from a real data distribution x_0, Gaussian noise is added multiple times until a noise sample x_T from a standard normal distribution is obtained. A reverse process then progressively removes this noise to recover the original data sample x_0. By introducing diffusion models to the remote sensing image generation domain, initial remote sensing images can be generated.

1) Forward Process (Noising Process)

For an original remote sensing image $x_0 \sim q(x_0)$, the forward diffusion process progressively adds Gaussian noise at each step, generating a series of noisy image samples $x_1,...,x_T$. Each step adds Gaussian noise to the data x_{t-1} from the previous step as follows:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{2}$$

Here, I is the identity matrix, t is the time step, and β_t is the variance used at each step, valued between 0 and 1. Typically, later steps use a larger variance, satisfying $\beta_1 < \beta_2 < \ldots < \beta_T$. Under a well-designed variance schedule, as $T \to \infty$, the final x_t completely loses the original information and becomes random noise:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{3}$$

Here, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\alpha_t = 1-\beta_t$ is the proportion of the original data retained at each step.

2) Reverse Process (Denoising Process)

The reverse process attempts to learn a denoising model $p_\theta(x_{t-1} \mid x_t)$ to recover the real data x_0 from the noisy data x_T. Similar to the forward process, the reverse process can be defined as a Markov chain:

$$p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t) \tag{4}$$

Therefore, starting from random noise $x_T \sim \mathcal{N}(0, I)$ and progressively denoising it can generate a real sample. The reverse process is thus the remote sensing image generation process. Where $p(x_T) = \mathcal{N}(x_T; 0, I)$, and $p_\theta(x_{t-1} \mid x_t)$ is a parameterized Gaussian distribution:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{5}$$

The key to the model is parameterizing the mean μ_θ. By deriving $q(x_{t-1} \mid x_t, x_0)$ using Bayes' theorem and matching it with p_θ, it can be proven that the optimal form of μ_θ depends on an estimation of x_0 (i.e., $\hat{x}_0$). In DDPM (Ho et al., 2020), μ_θ is not predicted directly. Instead, a neural network $\epsilon_\theta(x_t, t)$ is trained to predict the noise ϵ added at time t. Given $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, μ_θ can be derived from ϵ_θ:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) \tag{6}$$

Here, the variance $\Sigma_\theta(x_t, t)$ is typically fixed as $\tilde{\beta}_t I$, where $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

3) Training Objective

The training objective can be expressed as the L2 loss between ϵ_θ and the true noise ϵ:

$$L_{simple}(\theta) = \mathbb{E}_{x_0, \epsilon\sim\mathcal{N}(0,I), t}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2\right] \tag{7}$$

3.2.2 Hierarchical Caption Encoding for Remote Sensing Scene Generation: As illustrated in Figure 2, the hierarchical caption is mapped into segmented and layered conditional signals. For this segmented and layered conditional encoding, our study designs a temporally scheduled conditional signal generation mechanism. First, the hierarchical caption $C_{hierarchical}$ is obtained via the following expression:

$$C_{hierarchical} = \{C_{scene},\ C_{relation},\ C_{element}\} \quad (8)$$

It is then encoded into a feature representation E_c using a pre-trained language model:

$$E_c = Encoder(C_{hierarchical}) \quad (9)$$

Here, the scene, relation, and element layers correspond to different semantic granularities. A conditional fusion module injects the weighted hierarchical features into the U-Net's cross-attention mechanism, achieving progressive scene generation from macroscopic layout control to microscopic detail enhancement.

3.2.3 Knowledge-Aware Diffusion Model for Fine-Grained Remote Sensing Scene Generation: As shown in Figure 2, to accelerate the generation efficiency of remote sensing images, we adopt the approach of Stable Diffusion, centering on latent space diffusion. An encoder ε is used to compress the image into a low-dimensional latent data z_0, whose dimensionality is 48 times smaller than the original image space:

$$z_0 = \varepsilon(x_0) \quad (10)$$

After the diffusion process is completed in the latent space, a decoder $\mathcal{D}$ is used to decode the resulting latent data back into an image. This greatly reduces computational complexity and memory consumption while preserving the image's important features:

$$x_0' = \mathcal{D}(z_0') \quad (11)$$

The true power of Stable Diffusion lies in its ability to accept conditional inputs, enabling image generation from text prompts. This is achieved by converting the internal diffusion model into a conditional image generator using a cross-attention mechanism in the U-Net. For text inputs, they are first converted into embeddings (vectors) using a language model τ_θ (e.g., CLIP), and then they are mapped to the intermediate layers of the U-Net via a cross-attention layer:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V \quad (12)$$

For other spatially aligned inputs (e.g. semantic maps, images, inpainting), the conditioning can be done using concatenation. By inputting the hierarchical caption encoding E_c into the diffusion model, progressive denoising synthesis from semantics to image is achieved using text-conditional cross-attention. Finally, to further enhance training efficiency, various efficient fine-tuning algorithms are employed. Taking LoRA fine-tuning as an example (Hu et al., 2021):

$$W = W_0 + \Delta W = W_0 + BA \quad (13)$$

Here, $A \in \mathbb{R}^{r\times d}$, $B \in \mathbb{R}^{d\times r}$, and $r \ll d$.

4. Experimental Setup and Data

4.1 Datasets

We selected the AAAAD (Adopt-Amend-Annihilate-Add Dataset) as the experimental dataset (Ren et al., 2025). This dataset comprises 11 common remote sensing scenes (church, commercial, dense residential, industrial, medium residential, park, railway station, resort, school, sparse residential, and square). The image resolution is 512×512. It contains 3510 high-quality hierarchical remote sensing image captions, totaling 154826 words and 7141 sentences. The average caption length is 44.11 words (approx. 279.27 characters), with the longest caption containing 104 words (approx. 679 characters). Captions feature a maximum of 6 sentences, an average of 2.03 sentences, and 3.43 spatial relations, enabling detailed captions of the positional distribution of key geographical elements within the remote sensing images.

4.2 Relevant Algorithms

The algorithms involved in this study primarily encompass image-to-text generation, text-to-image generation, and their corresponding model fine-tuning algorithms. Specifically, the image-to-text algorithms include Deep Danbooru (Liu et al., 2023) for geographical element label generation, and BLIP-2 (Li et al., 2023) for spatial relation generation. The text-to-image algorithms include the mainstream Stable Diffusion and RS-SD (Zhang et al., 2024), which was trained on a massive remote sensing image dataset (approx. 1 million training samples). The model fine-tuning algorithms include current mainstream methods: Fine-Tune, Textual Inversion (Gal et al., 2022), LoRA, and DreamBooth (Ruiz et al., 2023).

4.3 Evaluation Metrics

The purpose of image quality assessment is to quantitatively compare the differences between generated and real images, thereby reflecting the generative model's performance, which primarily includes image quality and diversity. To comprehensively measure the quality of the generated images, we selected full-reference metrics such as PieAPP (Prashnani et al., 2018), LPIPS (Zhang et al., 2018), and GSHPS (Ren et al., 2024), as well as no-reference metrics including FID (Heusel et al., 2017), Q-Align (Wu et al., 2023), and CLIPScore (Hessel et al., 2021). This selection allows for a thorough assessment from different perspectives, aiding researchers in fully understanding and improving the performance of scene generation methods.

4.4 Model Implementation Details

We utilized the Stable Diffusion v1.5 pre-trained model as the foundation and employed Fine-Tune, Textual Inversion, DreamBooth, and LoRA algorithms to train the remote sensing scene generation models on the hierarchical captions. It is important to note that since RS-SD was already trained on a massive volume of remote sensing images (approx. 1 million training samples), which includes images from AAAAD, it can be directly applied to remote sensing scene generation without requiring additional training.

The dataset was split into training and test sets at an 8:2 ratio. During the training phase, the optimizer was AdamW8bit, and the initial learning rate was set to 1e-4 with a "cosine" decay strategy. The batch size was set to 10, and the maximum input token length was 75. The image resolution was 512×512, and mixed-precision training (fp16) was adopted. Each model was trained for 10 epochs to ensure an efficient training process and resource utilization. During the testing phase, the generated image resolution was consistent with the training phase (512×512). We used DDIM sampling for 30 steps, a CFG set to 7, and a Denoising strength set to 0.7. All experiments in this study were conducted on a single NVIDIA GeForce RTX 4090 and a single Quadro RTX 6000, using the Pytorch 2.1.0 deep learning framework and CUDA 12.1.

5. Results and Discussion

To systematically evaluate model performance, we utilized hierarchical captions to fine-tune multiple different text-to-image algorithms, obtaining several remote sensing image generation models. We then used different levels of captions as prompts to generate remote sensing images. These results were

compared against current state-of-the-art (SOTA) image generation models to evaluate the difference in generated image quality between hierarchical captions and traditional single-level captions.

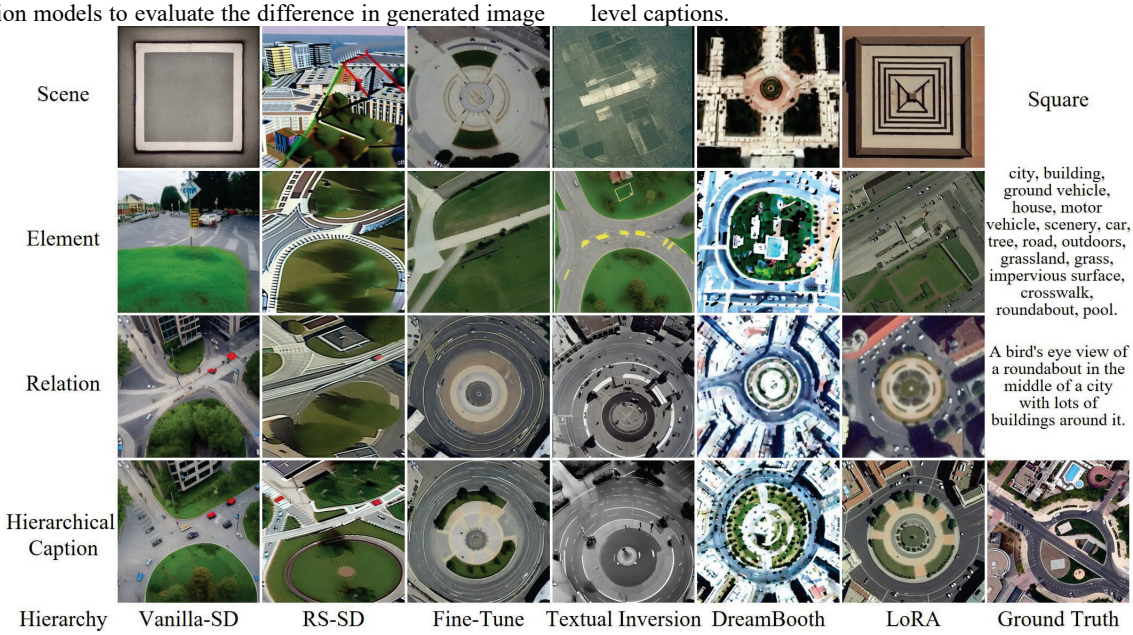

Figure 3. Direct comparison of generated images using single-level captions and hierarchical captions as prompts.

Method	Model	Metric					
		Full Reference			No Reference		
		PieAPP↓	LPIPS↓	GSHPS↑	FID↓	Q-Align↑	CLIPScore↑
Vanilla-SD	Traditional	2.87	**0.81**	**0.66**	**268.20**	4.11	**0.77**
	Hierarchical	**2.73**	0.82	0.64	279.89	**4.25**	**0.77**
RS-SD	Traditional	**2.70**	**0.77**	**0.63**	321.63	2.63	0.69
	Hierarchical	2.75	**0.77**	**0.63**	**319.42**	**2.67**	**0.71**

Table 1. Direct input of traditional single-level captions and hierarchical captions into existing text-to-image models. "Traditional" represents traditional single-level caption, and "Hierarchical" represents hierarchical caption. Bold and underlined values indicate the optimal result.

Method	Model	Metric					
		Full Reference			No Reference		
		PieAPP↓	LPIPS↓	GSHPS↑	FID↓	Q-Align↑	CLIPScore↑
Fine-Tune	Traditional	3.08	0.70	0.84	**240.28**	**3.06**	**0.74**
	Hierarchical	**2.90**	**0.69**	**0.88**	242.57	3.04	**0.74**
Textual Inversion	Traditional	3.38	**0.78**	0.61	274.33	**3.46**	**0.72**
	Hierarchical	**3.20**	0.79	**0.63**	**268.59**	**3.46**	0.70
DreamBooth	Traditional	3.26	0.72	0.82	213.59	1.66	0.73
	Hierarchical	**2.80**	**0.68**	**0.88**	**202.44**	**1.84**	**0.76**
LoRA	Traditional	3.14	0.70	0.86	228.43	2.46	0.72
	Hierarchical	**2.51**	**0.67**	**0.92**	**205.59**	**2.81**	**0.75**

Table 2. Fine-tuning existing algorithms with hierarchical captions, and inputting traditional single-level captions and hierarchical captions into the fine-tuned text-to-image models. "Traditional" represents traditional single-level caption, and "Hierarchical" represents hierarchical caption. Bold and underlined values indicate the optimal result.

Figure 3 clearly demonstrates the significant improvement in generation quality afforded by hierarchical captions. Starting with the scene concept, a concise word like "*square*" can quickly generate the basic annular structure of a square but often lacks detail and surroundings, as seen in the result from the Fine-Tune model (row 1, column 3). Furthermore, results like those in row 1, columns 1 and 6, still exhibit issues with ambiguity. The introduction of geographical elements (e.g.,

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-253-2026 | © Author(s) 2026. CC BY 4.0 License.

"*cars, trees, buildings*") can enrich the scene content, but may result in unreasonable spatial layouts. The spatial relation caption further enhances the structural accuracy of the image. For instance, "*A bird's eye view of a roundabout in the middle of a city with lots of buildings around it*" significantly improves the spatial structural accuracy, making the model aware that the scene to be generated is not just a "*square*" but one that must also contain a "*roundabout*". The stark difference between the generated results from the second and third rows clearly illustrates this conclusion. The hierarchical captioning method, by integrating these three levels, not only preserves the overall scene structure but also adds rich details and an accurate spatial layout. This is fully embodied in the results from the LoRA model in the final row, where we can observe a complete square scene that is detail-rich, geospatially sound, and contains a large roundabout with surrounding buildings and greenery. Hierarchical captions combine the global perspective of scene concepts, the rich detail of geographical elements, and the precise layout of spatial relations, further enhancing the realism and complexity of the generated image. This descriptive approach effectively overcomes the limitations of single-level captions, significantly improving the quality and diversity of remote sensing image generation and offering a more comprehensive and effective solution for the task.

To validate the effectiveness of the hierarchical captions, we conducted a series of comparative experiments. First, as shown in Table 1, we input hierarchical captions and traditional single-level captions into text-to-image models without any fine-tuning. The experimental results exposed a critical issue: when using the traditional Vanilla-SD model, the hierarchical captions performed worse than traditional captions, achieving optimal results on only 3 of the 6 metrics, while traditional captions secured 4. Our method was at a disadvantage, particularly on the key FID metric (279.89 vs. 268.20). This strongly indicates that for a model not specifically trained for it, our more complex, information-rich, structured caption may instead become a "burden", as the model cannot effectively parse its hierarchical information. Interestingly, when using the RS-SD model, which already possesses remote sensing domain knowledge, the situation improved; hierarchical captions achieved optimal results on 5 of the 6 metrics. This suggests that the model's prior knowledge of the domain aids in understanding our structured information.

As shown in Table 2, we fine-tuned various text-to-image algorithms using the hierarchical captions. The results further confirmed our hypothesis: not all fine-tuning methods can effectively utilize these structured, hierarchical captions. After standard Fine-Tune and Textual Inversion fine-tuning, the performance of hierarchical captions remained unstable. For example, with Fine-Tune, traditional captions still held an advantage in FID (240.28 vs. 242.57) and Q-Align (3.06 vs. 3.04). With Textual Inversion, traditional captions won on 4 of the 6 metrics. This series of negative results highlights a limitation of our method: it places higher demands on the decoding capability of the text-to-image model's text encoder. However, when fine-tuning with DreamBooth and LoRA, the potential of hierarchical captions was fully unleashed, achieving optimal results on all six evaluation metrics. Using LoRA as an example, hierarchical captions substantially optimized PieAPP from 3.14 to 2.51, increased GSHPS from 0.86 to 0.92, and significantly reduced FID from 228.43 to 205.59. These results provide strong evidence that once a text-to-image model learns to understand this structured, hierarchical input through fine-tuning, it can guide the generation process more precisely and comprehensively than traditional single-level captions, thereby significantly enhancing the quality and fidelity of the generated images.

6. Conclusion

This paper addresses the problem of insufficient control precision in fine-grained remote sensing image generation with diffusion models. We propose a fine-grained remote sensing image generation method driven by expert knowledge and hierarchical captions. This method deconstructs remote sensing images into a hierarchical "element-relation-scene" structure and injects this structure into the diffusion model's U-Net via a novel encoding mechanism, significantly improving the fineness and controllability of the generated results.

Our experimental results reveal a key finding: the effectiveness of hierarchical captions is highly correlated with the fine-tuning strategy employed. The key to unlocking the potential of these structured captions lies in the synergistic adaptation between the model and the captions. The study demonstrates that base models without specific fine-tuning (such as Vanilla-SD or the domain-aware RS-SD) struggle to effectively parse this complex structured information, with generation results that are sometimes inferior to those from traditional single-level captions. However, when our method is combined with efficient fine-tuning algorithms like LoRA or DreamBooth, its performance is fully realized, achieving optimal results across all six evaluation metrics. This strongly proves that after specific training, a diffusion model can learn to decouple and utilize this hierarchical semantic guidance, achieving precise control from macroscopic layout to microscopic detail.

Despite its success, this work also reveals the method's dependency on specific fine-tuning techniques (e.g., LoRA). Future research should explore more robust conditional injection architectures to reduce this dependency and to further enhance the accuracy and robustness of the automated hierarchical caption generation pipeline. Furthermore, applying this hierarchical control framework to generation tasks in other specialized domains, such as medical imaging or architectural design, represents a valuable direction for future exploration.

Acknowledgements

This study was funded by the Major Program of the National Natural Science Foundation of China under Grants 42394060 and 42394062, the National Key Research and Development Program of China under Grant 2022YFB3904205, and the Chongqing Municipal Planning and Natural Resources Bureau 2025 Annual Scientific Research Program Major Project under Grant KJ2025027.

References

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D., 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. *arXiv preprint arXiv:2208.01618*.

Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y., 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *2021 Conference on empirical methods in natural language processing (EMNLP 2021)*: 7514-7528.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840-6851.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401-4410.

Li, J., Li, D., Savarese, S., Hoi, S., 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, International conference on machine learning. *Proceedings of machine learning research*, pp. 20351-20383.

Liu, W., Chen, H., Ren, J., Zhang, Z., Li, R., Zhao, T., Zhai, X., Zhu, X., 2024. Research on knowledge extraction from street scene images based on hybrid intelligence. *Acta Geodaetica et Cartographica Sinica*, 53(09): 1817-1828.

Liu, W., Chen, J., Ren, J., Xu, C., Li, R., Zhai, X., Jiang, Z., Zhang, Y., Peng, Y., Wang, X., 2022. Hybrid Intelligence-Based Framework for Automatic Map Inspecting Technology. *Geomatics and Information Science of Wuhan University,* 47(12): 2038-2046.

Liu, Y., Yu, C., Shang, L., Wu, Z., Wang, X., Zhao, Y., Zhu, L., Cheng, C., Chen, W., Xu, C., 2023. FaceChain: A Playground for Identity-Preserving Portrait Generation. *arXiv preprint arXiv:2308.14256.*

Lu, X., Wang, B., Zheng, X., Li, X., 2018. Exploring models and data for remote sensing image caption generation. *IEEE Trans. Geosci. Remote Sens.*, 56(4): 2183-2195.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784.*

Prashnani, E., Cai, H., Mostofi, Y., Sen, P., 2018. Pieapp: Perceptual image-error assessment through pairwise preference, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1808-1817.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ren, J., Liu, W., Chen, J., Li, Z., Yin, S., Zhang, J., 2025. Key challenges and countermeasures for automated recognition of problematic maps. *Journal of Spatio-temporal Information*, 32(6): 606-619.

Ren, J., Liu, W., Chen, J., Yin, S., 2025. HI4HC and AAAAD: Exploring a hierarchical method and dataset using hybrid intelligence for remote sensing scene captioning. *Int. J. Appl. Earth Obs. Geoinf.,* 139: 104491.

Ren, J., Liu, W., Chen, J., Yin, S., Tao, Y., 2024. Word2Scene: Efficient remote sensing image scene generation with only one word via hybrid intelligence and low-rank representation. *ISPRS J. Photogramm. Remote Sens.*, 218: 231-257.

Ren, J., Liu, W., Chen, J., Zhang, J., Yin, S., 2026. SAFE: sensitive annotation finding and extraction from multi-type Chinese maps via hybrid intelligence and knowledge graph. *Geo-Spat. Inf. Sci.*: 1-26.

Ren, J., Liu, W., Chen, J., Zhang, L., Tao, Y., Zhu, X., Zhao, T., Li, R., Zhai, X., Wang, H., Zhou, X., Hou, D., Wang, Y., 2024. Knowledge-guided intelligent recognition of the scale for fragmented raster topographic maps. *Acta Geodaetica et Cartographica Sinica*, 53(01): 146-157.

Ren, J., Liu, W., Li, Z., Li, R., Zhai, X., 2021. Intelligent Detection of "Problematic Map" Using Convolutional Neural Network. *Geomatics and Information Science of Wuhan University*, 46(4): 570-577.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684-10695.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K., 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500-22510.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, *International conference on machine learning*, pp. 6105-6114.

Tang, D., Cao, X., Hou, X., Jiang, Z., Liu, J., Meng, D., 2024. CRS-Diff: Controllable remote sensing image generation with diffusion model. *IEEE Trans. Geosci. Remote Sens.*, 62: 1-14.

Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., 2023. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090.*

Yin, S., Liu, W., Chen, J., Ren, J., Tao, Y., Wang, Y., Zhang, J., 2025. HIUFE: Hybrid intelligence-based unauthorized farmland excavation scene cognition. *ISPRS J. Photogramm. Remote Sens.*, 227: 276-296.

Yin, S., Liu, W., Chen, J., Ren, J., Zhang, J., 2026. TopoFarm: A Topology-Annotated Panoptic Dataset for Unauthorized Farmland Excavation Scene Representation. *ISPRS Int. J. Geo-Inf.*, 15(3): 93.

Zhang, J., Chen, J., Fan, H., Zhou, X., Hou, D., Ren, J., Yin, S., Hou, M., 2025. Shp2gml: semantic 3D model generation for Ming and Qing historical buildings at multiple LoDs using domain knowledge and multi-source data. *Int. J. Digit. Earth*, 18(2): 2564910.

Zhang, J., Hou, M., Chen, J., Zhou, X., Hou, D., Ren, J., Yin, S., Zhao, H., Zhang, Z., 2026. A CityGML ADE for modeling ancient chinese timber architecture in 3D with semantic information. *npj Heritage Science*, 14(1): 271.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586-595.

Zhang, Z., Zhao, T., Guo, Y., Yin, J., 2024. RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Trans. Geosci. Remote Sens.*, 62: 1-23.