Landslide Retrieval and Reranking for Controllable Diffusion Generation of Post-Disaster UAV Imagery

Yiyang Ma[a], Han Hu[a,*], Guozhong Lin[a], Lin Chen[b], Junfan Wang[a], Yulin Ding[a], Yang Jia[c], Bo Xiang[c], Jiwei Deng[d] and Qing Zhu[a]

[a]Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu, 611756, China
[b]POWERCHINA Chengdu Engineering Corporation Limited, Chengdu, China
[c]Sichuan Highway Planning, Survey, Design and Research Institute Ltd., Chengdu, 610041, China
[d]China Railway Design Corporation, Tianjin, 300251, China

ARTICLE INFO

Keywords:
Image Generation
diffusion models
deep learning
Remote Sensing Image

ABSTRACT

Post-disaster unmanned aerial vehicle (UAV) imagery of landslides is scarce because acquisition is constrained by narrow response windows. Generative models offer a new way to expand the available samples. To generate landslide images with the desired appearance and spatial layout, these models need to combine textual descriptions, visual styles, and semantic masks. To address this challenge, we propose LRRGen, a retrieval-augmented framework for controllable UAV landslide image synthesis. Unlike methods that encode the reference image as a whole for appearance guidance, LRRGen directly retrieves real images from a reference database and injects their local visual features into corresponding semantic regions during generation. First, we construct a controlled vocabulary of landslide visual attributes, organized as a Visual Style Taxonomy, to standardize scene descriptions and establish an aligned image–semantic-mask–text reference database. Second, we introduce a two-stage reference selection strategy that combines coarse multimodal retrieval with fine-grained reranking, incorporating region-area compatibility to match references to the target layout. Finally, we adopt Attention Distillation with semantic matching to transfer reference textures within corresponding semantic regions, while text conditioning and ControlNet jointly guide scene content and spatial structure. Experiments show that LRRGen achieves a landslide-region IoU of 0.65, compared with 0.62 for Vanilla ControlNet and 0.57 for IP-Adapter + ControlNet, together with the best distributional metrics in both SigLIP and DINOv2 feature spaces and the highest overall VQA score of 6.23 among the evaluated generation methods. Further analyses demonstrate controllable variation in landslide color and texture and improved integration of reference appearance within the prescribed semantic layout.

1. INTRODUCTION

Post-disaster unmanned aerial vehicle (UAV) images support disaster detection and damage assessment (Rahnemoonfar et al., 2021; Li et al., 2024), as well as emergency response (Kyrkou and Theocharides, 2020). However, images that show landslides or debris flows damaging roads and buildings are rare. Such images are difficult to collect because emergency road-clearance operations can rapidly alter the scene; for example, a national trunk-road disruption requiring 12 hours or more to resolve is classified as a major emergency (MOT, 2010). Old-landslide imagery cannot replace post-event imagery of event-triggered landslides (Gao et al., 2024; Lee et al., 2008). Over time, vegetation recovery, erosion, and human activity alter the visual features of old landslides (Gao et al., 2024). By contrast, post-event images capture fresh slope failures caused by a specific disaster (Lee et al., 2008). Old-landslide datasets also rarely show direct damage to roads, buildings, and other assets related to human safety and property loss. A model trained mainly on old-landslide imagery may therefore perform poorly in real emergency surveys.

Diffusion models provide a practical way to fill this data gap (Ho et al., 2020; Song et al., 2021b; Rombach et al., 2022a). They can generate diverse post-disaster UAV images that include both hazards and nearby infrastructure. Following a common synthetic-to-real training strategy, these images can be used to pre-train downstream models, which are then fine-tuned with a small set of real post-disaster images to improve their reliability and coverage (Zheng et al., 2023; Taskina et al., 2026). However, synthetic images are useful for downstream training only when the generation process is controllable: hazards and exposed infrastructure must follow a prescribed spatial layout, while their visual appearance and style must remain realistic. Semantic-map guidance is often used to determine the locations, extents, and boundaries of landslide, road, and background regions (Zhang et al., 2023); textual conditioning specifies scene semantics and region-level visual attributes (Rombach et al., 2022a); and reference-image conditioning supplies fine-grained textures and visual styles that are difficult to describe in language (Ye et al., 2023). Therefore, this work focuses on integrating these complementary forms of guidance—semantic maps, text, and reference images—to generate realistic post-disaster UAV imagery with prescribed spatial layouts and fine-grained local details.

In terms of conditioning, existing methods mainly follow two strategies, as illustrated in Figure 1(a) and (b). First, as shown in Figure 1(a), this strategy primarily uses textual descriptions to control visual style and variations in local

*Corresponding author
Email address: han.hu@swjtu.edu.cn (H. Hu)

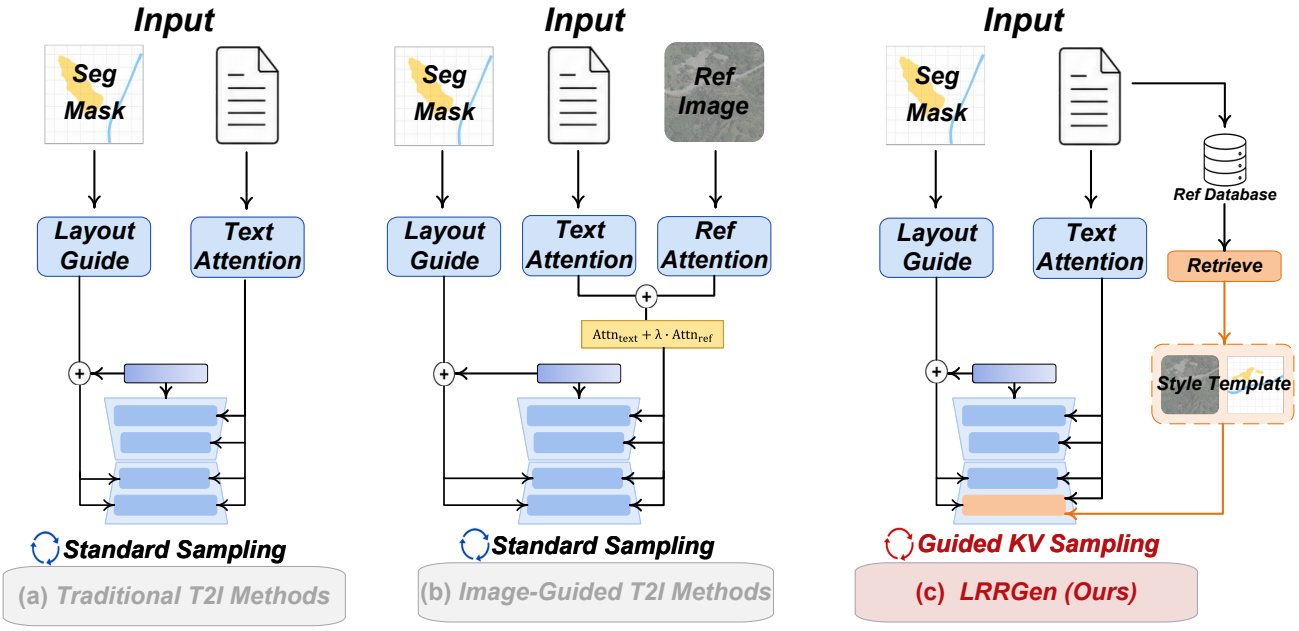


Figure 1: Comparison of three conditioning strategies for post-disaster UAV landslide image generation: (a) text and semantic-map guidance, (b) reference-image guidance, and (c) the proposed retrieval-augmented, region-specific guidance.

details, while semantic-map guidance constrains the scene layout (Rombach et al., 2022a; Zhang et al., 2023). Second, as shown in Figure 1(b), this strategy encodes visual features from a reference image and injects them into the generation process to control texture and visual style (Ye et al., 2023).

As shown in Figure 1(a), text and semantic maps provide only coarse control over appearance and layout, limiting the synthesis of fine-grained, region-specific textures. As shown in Figure 1(b), global reference-image guidance may transfer irrelevant textures across semantic regions and disrupt the prescribed layout. Both strategies rely on shared representations learned by general-purpose vision-language encoders to connect textual or reference-image conditions with the generation process. However, these encoders are primarily pretrained on natural image–text pairs and remain less reliable for UAV remote-sensing scenes, where overhead viewpoints, large scale variations, small objects, and fine-grained local textures weaken cross-modal alignment (Liu et al., 2024; Kuckreja et al., 2024). To address these limitations, this work adopts the strategy shown in Figure 1(c), which decouples spatial-structure control from region-level visual-style injection. Semantic masks constrain the locations, extents, and boundaries of landslide, road, and background regions, while reference images aligned with the target conditions are automatically retrieved from a real-image library (Chen et al., 2023; Shalev-Arkushin et al., 2026). The retrieved visual features are injected only into the corresponding semantic regions, preventing cross-region interference while preserving the global layout.

Building on the above principle, we propose LRRGen (Landslide Retrieval, Reranking, and Generation), a controllable image-generation framework built on Stable Diffusion (Rombach et al., 2022a) and implemented with the Hugging Face Diffusers library (von Platen et al., 2022). LRRGen retrieves and reranks real landslide images as visual priors for generating post-disaster UAV landslide imagery. To obtain real visual priors that are well aligned with the target scene, we first construct a Visual Style Taxonomy that organizes region-specific visual attributes of landslide and road regions into structured style descriptions. Before generation, LRRGen selects a reference image from the real-image library through coarse-to-fine retrieval and reranking. Qwen3-VL-Embedding first retrieves candidates based on the target description, and Qwen3-VL-Reranker refines their ranking; a region-area-ratio constraint further ensures compatibility with the target layout (Li et al., 2026). During diffusion sampling, semantic conditioning maps constrain the locations, extents, and boundaries of individual regions, while the retrieved references provide region-specific texture priors (Zhou et al., 2025). These priors are transferred through Guided KV Sampling, which selectively incorporates semantically matched reference key–value features into the attention computation instead of directly injecting global reference features into the denoising network.

In summary, the main contributions of this work are: (1) a Visual Style Taxonomy for post-disaster UAV landslide imagery that structures region-level visual attributes for consistent description, annotation, and retrieval of real visual priors; and (2) LRRGen, a retrieval-augmented controllable

generation strategy that automatically retrieves matched reference images and transfers their texture priors to corresponding semantic regions while preserving the target spatial layout.

2. RELATED WORK

The following review focuses on two areas: text-to-image generation and visual style control, with particular emphasis on their applications to remote sensing imagery.

2.1. Text-to-Image Generation for Remote Sensing

The development of generative models in computer vision has undergone a paradigm shift from Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) to Diffusion Models (Ho et al., 2020; Song et al., 2021b). Since LRRGen is built on diffusion models, the following discussion focuses on diffusion-based text-to-image generation. Diffusion models have become a leading paradigm in image synthesis owing to their strong distribution-modeling capacity and stable training. DALL-E 2 (Ramesh et al., 2022) strengthened text-guided generation through CLIP-based cross-modal alignment (Radford et al., 2021), while GLIDE (Nichol et al., 2022) explored different guidance strategies. Latent Diffusion Models (Rombach et al., 2022a), which form the basis of Stable Diffusion (Rombach et al., 2022b), further reduced the computational cost of high-resolution image generation by performing denoising in a compact latent space.

With the success of diffusion models in natural image generation, researchers have begun to introduce them into remote sensing image synthesis to support data augmentation and downstream perception tasks. Existing diffusion-based studies can be broadly grouped into three directions.

First, conditional generation uses text or numerical metadata to control scene content and geographic attributes. DiffusionSat (Khanna et al., 2024) injects coordinates and ground sample distance, whereas RSDiff (Sebaq and El-Helw, 2024) and Text2Earth (Liu et al., 2025) learn text-conditioned remote sensing distributions; Text2Earth further scales this paradigm to global, multi-resolution data. Second, spatially controlled generation introduces explicit structural conditions to improve layout consistency. SatSynth (Toker et al., 2024) synthesizes paired aerial images and semantic masks, while CRS-Diff (Tang et al., 2024) jointly models text, metadata, semantic masks, road maps, and sketches. Third, cross-scale generation addresses resolution variation and spatial continuity. ZoomLDM (Yellapragada et al., 2025) supports multi-scale generation, while MetaEarth (Yu et al., 2025) uses a resolution-guided self-cascaded framework to synthesize continuous, unbounded remote sensing scenes.

Despite this progress, two challenges remain particularly relevant to text-conditioned remote sensing generation. First, reliable image–text alignment remains difficult because general-purpose vision–language encoders are trained mainly on natural images and do not fully capture overhead viewpoints, scale variation, and domain-specific remote sensing semantics (Liu et al., 2024). Second, even with adequate semantic alignment, text prompts cannot fully specify continuous, fine-grained visual styles such as material composition, surface roughness, and local texture (Duan et al., 2025). Therefore, text-only conditioning is insufficient for our task, motivating the use of explicit spatial constraints and reference-image guidance.

2.2. Visual Style Control for Remote Sensing Image Generation

Beyond text, controllable diffusion methods commonly use two complementary forms of visual guidance: structured condition maps for geometry and reference images for appearance. ControlNet and Uni-ControlNet inject edges, depth maps, semantic masks, or multiple local and global conditions through auxiliary control branches, thereby constraining spatial layout and object geometry (Zhang et al., 2023; Zhao et al., 2023). Reference-guided methods instead transfer texture and style from visual examples: IP-Adapter injects global image embeddings through decoupled cross-attention, whereas StyleID transfers reference key–value features while retaining the target queries to preserve content structure (Ye et al., 2023; Chung et al., 2024). Together, these approaches provide explicit control over the geometric structure and visual appearance of generated images.

In remote sensing image synthesis, controllable generation has similarly progressed from text-only prompts toward richer conditioning signals, including geographic metadata, semantic masks, road maps, sketches, and reference images. For example, GeoSynth (Sastry et al., 2024) uses map-derived layouts and geographic cues, whereas CRS-Diff (Tang et al., 2024) integrates text, metadata, semantic masks, road maps, and sketches through multi-scale fusion. EarthSynth (Pan et al., 2025) and Any2RSI (Zhang et al., 2026) combine textual semantics with semantic masks or multiple spatial controls to improve layout consistency and semantic coherence. TISynth (Dong et al., 2025) further incorporates real reference images with semantic masks and text prompts, extending control from spatial structure to visual appearance.

At the fusion level, heterogeneous conditions are typically encoded by modality-specific branches and incorporated into the denoising network through multi-scale feature fusion, residual connections, or attention mechanisms. CRS-Diff employs multi-scale fusion, while Any2RSI introduces a cross-modal multi-control adapter to align textual semantics with multiple spatial signals (Tang et al., 2024; Zhang et al., 2026). For reference-aware synthesis, TISynth uses reference-semantic joint learning and an All-in-Attention module to fuse reference images, semantic masks, and text prompts within the attention layers (Dong et al., 2025).

Most existing methods operate at the global level, controlling scene layout, regional attributes, or overall appearance, and therefore cannot adequately model region-specific local details. These local landslide characteristics, such as

color, texture, debris patterns, and surface morphology, are difficult to describe explicitly and are therefore more effectively captured from real reference imagery.

2.3. Landslide Visual Style Taxonomy

To improve prompt controllability, we organize landslide appearances into structured visual categories, enabling relevant attributes to be extracted and combined into consistent prompts. Landslides are conventionally classified by the displaced material and the type of movement. The updated Varnes system combines material categories such as rock, debris, and earth with movement types such as fall, topple, slide, spread, and flow (Hungr et al., 2014). This process-based taxonomy is essential for geological interpretation, but it does not directly describe how a landslide appears in an optical image; landslides assigned to the same type may still differ substantially in color, texture, surface roughness, and depositional pattern.

Image-based interpretation therefore characterizes landslides through a combination of visual cues. Common criteria include shape, size, color or tone, texture, spatial pattern, topographic position, and surrounding context (Guzzetti et al., 2012). Martha et al. (Martha et al., 2010) further combined spectral, spatial, contextual, and morphological properties to distinguish debris slides, debris flows, and rock slides. At the finer scale of UAV imagery, landslide appearance can be described by color or spectral response, texture, and shape, with exposed materials, cracks, and local surface structures providing further detail (Cheng et al., 2025).

These classification and interpretation studies provide a useful basis for describing landslide appearance. Drawing on their visual criteria, we use color, texture, and morphology to construct structured prompts for controllable generation.

3. Method

3.1. Framework Overview

As illustrated in the upper part of Figure 2, LRRGen is built upon the widely used Stable Diffusion v1.5 architecture (Rombach et al., 2022a,b) and employs ControlNet (Zhang et al., 2023) for semantic-mask-based spatial control. Although CLIP-based text conditioning provides effective semantic guidance, the CLIP text encoder used in Stable Diffusion v1.5 is a relatively lightweight model trained primarily for global image–text alignment rather than fine-grained visual reasoning (Radford et al., 2021). Its compact text representation tends to compress subtle appearance cues, such as soil granularity, debris patterns, erosion traces, and material transitions, into coarse semantic concepts. Consequently, natural-language prompts alone have limited capacity to specify the fine-grained, photorealistic textures of post-disaster landslide scenes. Inspired by reference-based style transfer and retrieval-augmented generation (Chung et al., 2024; Zhou et al., 2025; Chen et al., 2023; Shalev-Arkushin et al., 2026), LRRGen therefore retrieves a real image through a retrieval-and-reranking strategy and explicitly injects its region-aligned visual features into the diffusion model. The framework comprises the following four components.

ControlNet-based Stable Diffusion backbone. The generation backbone combines Stable Diffusion v1.5 with a ControlNet branch conditioned on the target semantic mask. The text prompt supplies scene semantics through CLIP-based cross-attention, while ControlNet encodes the mask into multi-scale spatial residuals that constrain the positions, extents, and boundaries of landslide, road, and background regions. This design provides a stable content-generation basis and preserves the prescribed spatial layout throughout diffusion sampling. During training, all other network components are kept frozen, and only the semantic-mask encoder of ControlNet is fine-tuned.

Landslide visual style taxonomy (Sec. 3.2). LRRGen constructs an image–semantic-mask–text reference database in which every sample comprises an aligned UAV image, semantic mask, and structured textual description. The UAV image provides authentic visual appearance, the semantic mask specifies the spatial layout of landslide, road, and background regions, and the structured text describes their visual attributes using the proposed taxonomy. By establishing explicit correspondences among visual content, semantic regions, and textual descriptions, this database provides standardized reference units for subsequent retrieval, reranking, and region-aware feature injection.

Sample retrieval and reranking (Sec. 3.3). LRRGen adopts the Qwen VLM as the retrieval backbone and combines its embedding and reranking capabilities in a cascaded two-stage mechanism (Li et al., 2026). Using the structured prompt as the sole query, Qwen3-VL-Embedding independently encodes the prompt and the image–text database entries for efficient full-gallery search and Top-K recall. Because these coarse similarities mainly capture global correspondence, they may confuse samples with similar semantics but different local appearances. Qwen3-VL-Reranker therefore jointly evaluates the prompt, candidate image, and textual annotation to model fine-grained cross-modal correspondence and reorder the recalled candidates.

Region-aware reference feature injection (Sec. 3.4). The retrieved sample is encoded by a dedicated reference branch and introduced into the diffusion U-Net through key–value feature injection. A semantic matching mask restricts feature transfer to corresponding landslide, road, and background regions, preventing cross-region style contamination. Together with text-based semantic guidance and ControlNet-based spatial constraints, this mechanism transfers realistic color distributions, debris textures, road materials, and local geomorphological patterns from real observations while preserving the target layout.

3.2. Visual Style Taxonomy for Landslide Imagery

Our previous work established a relatively large-scale semantic-segmentation dataset for UAV landslide imagery,

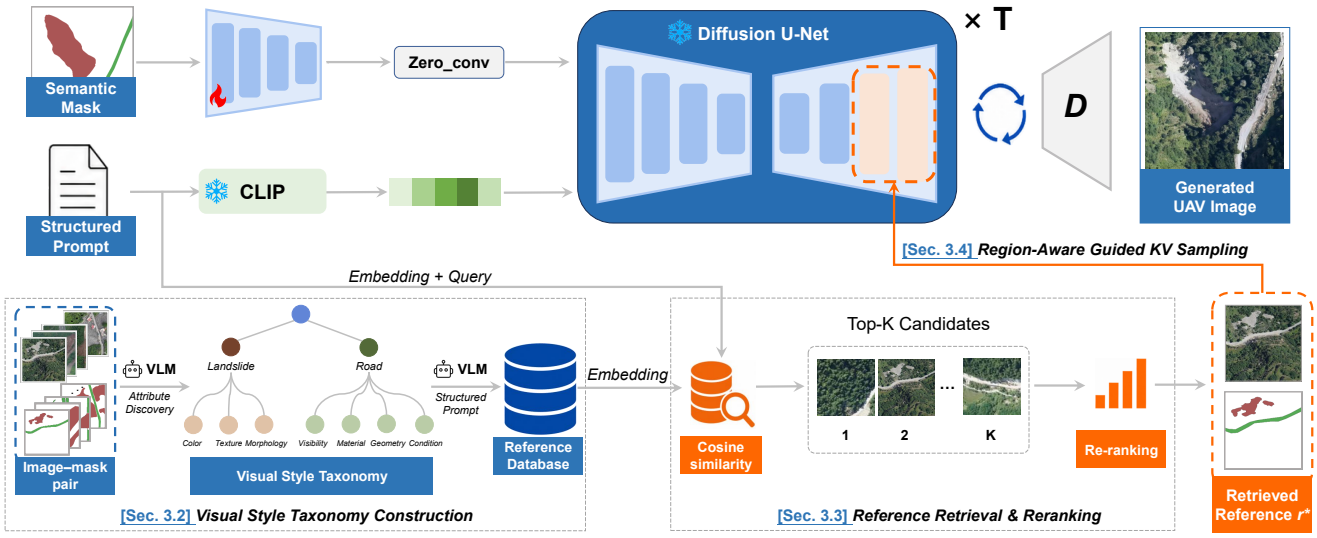


Figure 2: Overall framework of the proposed LRRGen for retrieval-augmented style-aware diffusion generation of UAV landslide imagery.

Table 1
Construction rules and representative textual descriptions for the landslide and road visual taxonomy.

Category	Attribute	Initial generation	Post-processing	Representative textual description
Landslide	Color	Segmented image + VLM (Gemini 3.1 Pro)	Manual adjustment	"light grey"
	Texture	Segmented image + VLM (Gemini 3.1 Pro)	Manual adjustment	"rough crumbled texture"
	Morphology	–	Manually specified	"tongue-shaped flow"
Road	Material	Segmented image + VLM (Gemini 3.1 Pro)	Manual adjustment	"grey asphalt"
	Geometry	Segmented image + VLM (Gemini 3.1 Pro)	Manual adjustment	"winding zig-zag"
	Surface condition	Segmented image + VLM (Gemini 3.1 Pro)	Manual adjustment	"aging surface"
	Visibility	–	Rule-based assignment	"dominant"

with pixel-level annotations of landslide and road regions (Gu et al., 2026). Building on this image–semantic foundation, the present work further introduces structured textual descriptions and proposes a Visual Style Taxonomy for constructing an *image–semantic–text* reference database of post-disaster UAV landslide imagery. This taxonomy decouples the naturally continuous, mixed, and hard-to-describe disaster visual appearances into a discrete set of visual attributes with clear semantic boundaries. Landslide regions are characterized by color, texture, and morphology, whereas road regions are characterized by material, geometry, surface condition, and visibility. Consequently, visual conditions in disaster imagery can be incorporated into subsequent reference-image retrieval and controllable generation in a standardized and composable form.

As shown in Figure 3, semantic masks are first used to construct region-guided visual inputs, from which initial visual attributes are extracted separately for the landslide and road regions. For landslide regions, we primarily extract landslide color from exposed soil and rock and a unified landslide texture description encompassing particle appearance, surface structure, and erosion characteristics. For road regions, the extracted attributes mainly include road material, road geometry, and road surface condition. The background region does not participate in controlled-vocabulary construction; instead, a constrained scene-context description is generated during the subsequent annotation stage.

To reduce interference from irrelevant visual context, we adopt a mask-guided regional visual prompting strategy (Yang et al., 2023; Shtedritski et al., 2023), which guides the multimodal vision-language model to focus on the designated region. Gemini 3.1 Pro Preview (Google, 2026) is then employed to generate initial region-level visual descriptions.

As summarized in Table 1, because open-ended descriptions are often mixed and ambiguous, candidate terms are

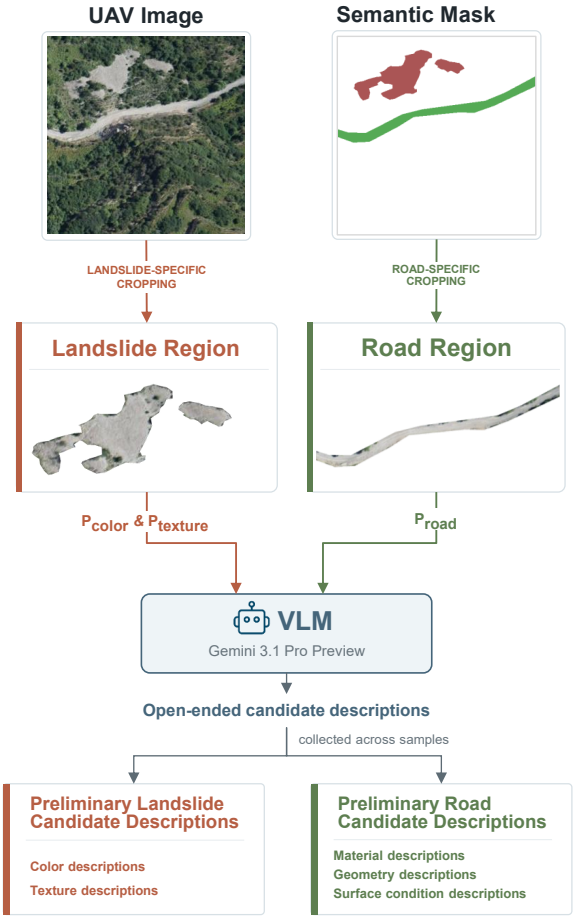


Figure 3: Region-guided and prompt-specific discovery of open-ended visual attributes. Semantic masks are first used to construct landslide- and road-specific visual inputs from post-disaster UAV imagery. Separate landslide color and landslide texture prompts are applied to the landslide view, whereas a unified road prompt jointly queries road material, road geometry, and road surface condition. The three prompts independently invoke a shared VLM, and the resulting open-ended descriptions are collected across samples to form preliminary landslide and road candidate pools.

manually consolidated. This manual consolidation process improves prompt controllability and ensures consistent descriptions across samples. The final taxonomy comprises seven attribute dimensions: landslide color, texture, and morphology, together with road material, geometry, surface condition, and visibility. Landslide color and texture and road material, geometry, and surface condition are obtained by consolidating Gemini-generated candidate descriptions. In contrast, landslide morphology is manually predefined, whereas road visibility is assigned using rules based on the road-mask area ratio.

3.3. Text-Prompt-Based Reference Retrieval and Reranking

Given a structured prompt constructed from the taxonomy above, LRRGen retrieves the most relevant reference sample from the database. Specifically, each database entry is indexed using a fused image–text representation; the prompt serves as the sole query for identifying the closest visual–textual match, after which the corresponding reference image and its paired semantic mask are retrieved together for subsequent generation.

Reference database embedding. Let $\mathcal{D} = \{(I_i, t_i, M_i)\}_{i=1}^{N}$ denote the reference database, where I_i, t_i, and M_i are the UAV image, its structured textual annotation, and the corresponding semantic mask, respectively. For offline indexing, Qwen3-VL-Embedding-8B (Li et al., 2026) jointly encodes I_i and t_i into a fused multimodal representation, which is ℓ_2-normalized as

$$g_i = \frac{E_{\mathrm{vl}}(I_i, t_i)}{\left\| E_{\mathrm{vl}}(I_i, t_i) \right\|_2}, \qquad g_i \in \mathbb{R}^C. \tag{1}$$

The embeddings of all gallery samples are stacked row-wise to form the index matrix $G = [g_1; \ldots; g_N] \in \mathbb{R}^{N \times C}$, where $N = 3{,}903$ in our reference database and $C = 4{,}096$ is the output dimension of Qwen3-VL-Embedding-8B. Each row of G retains the identifier of its paired image, annotation, and semantic mask, enabling the complete reference sample to be recovered after retrieval.

Prompt embedding and coarse retrieval. Given a structured query prompt t_q, the same model produces a text-only embedding $q = E_{\mathrm{vl}}(t_q)/\|E_{\mathrm{vl}}(t_q)\|_2 \in \mathbb{R}^C$. CLIP independently encodes an image and a text sequence and learns their global alignment through an image–text contrastive objective; consequently, the visual content and structured annotation of a gallery sample cannot interact within the encoder, and their fine-grained attribute correspondences may be compressed into separate global representations (Radford et al., 2021). In contrast, Qwen3-VL-Embedding processes the visual and textual tokens of a mixed-modal gallery entry within the same Qwen3-VL backbone and summarizes their fused hidden states as a single embedding. Its retrieval representation is further optimized through large-scale multimodal and multi-task contrastive learning, hard-negative mining, and distillation of fine-grained relevance scores from a cross-encoder reranker (Li et al., 2026). These mechanisms allow the database embedding to capture interactions between visual evidence and structured textual attributes before it is compared with the prompt, providing a more suitable cross-modal representation for the coarse retrieval stage in LRRGen.

Because both q and the rows of G are ℓ_2-normalized, all cosine similarities can be obtained in a single matrix–vector product:

$$\mathbf{s}^{\mathrm{coarse}} = Gq, \qquad s_i^{\mathrm{coarse}} = g_i^{\top} q = \cos(g_i, q). \tag{2}$$

We retain the $K = 10$ highest-scoring gallery samples as the coarse candidate set C_K, which is passed to the subsequent fine-grained reranking stage.

Tokenized and patch-wise fine-grained reranking. Although coarse retrieval efficiently searches the full database,

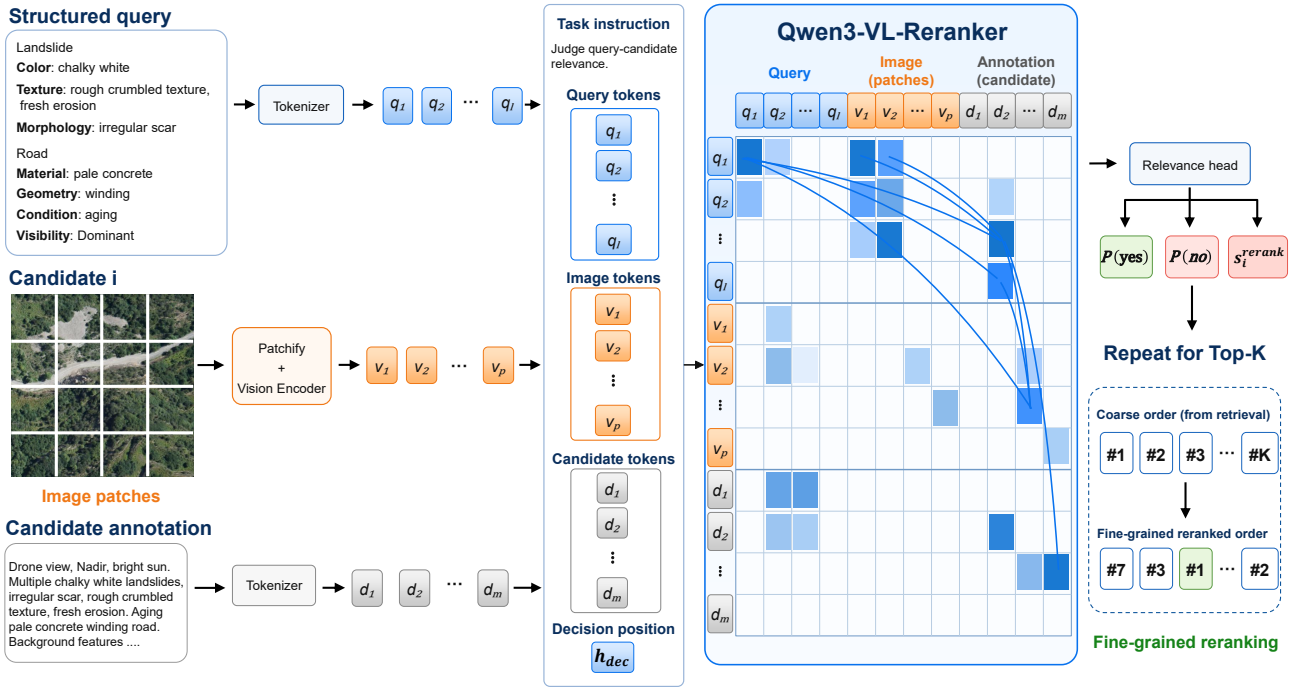


Figure 4: Token- and patch-level fine-grained reranking with Qwen3-VL-Reranker.

each query–candidate pair is compared only through a single global embedding. We therefore apply Qwen3-VL-Reranker-8B (Li et al., 2026) to the recalled candidate set $\mathcal{C}_K$, as illustrated in Figure 4. Specifically, the structured prompt is tokenized as $Q = \{q_\ell\}_{\ell=1}^{L}$, while the image of candidate i is divided into patches and encoded as visual tokens $V_i = \{v_{i,p}\}_{p=1}^{P_i}$; its annotation is likewise represented by textual tokens $D_i = \{d_{i,m}\}_{m=1}^{M_i}$. The Reranker processes Q, V_i, and D_i jointly, allowing individual query attributes to interact with localized image patches and the corresponding candidate annotation instead of being compressed into independently compared global vectors. A relevance head maps the resulting decision representation h_i^{dec} to a scalar score s_i^{rerank}, and the K candidates are sorted in descending score order to obtain the fine-grained ranking. This procedure is repeated independently for every candidate in $\mathcal{C}_K$ and does not require a fixed number of image patches.

Mask-based area-ratio weighting. Although the Reranker measures fine-grained semantic and appearance correspondence, it does not explicitly account for the relative spatial extent of the landslide region. LRRGen therefore introduces a soft area-consistency term computed directly from the input semantic mask M_q and the semantic mask M_i paired with candidate i. We denote by ρ_q and ρ_i the proportions of pixels labeled as landslide in M_q and M_i, respectively. Their agreement can be expressed as $a_i = 1 - |\rho_q - \rho_i|$, where a larger value indicates a more compatible landslide proportion. The Reranker scores in the current Top-K pool are first min–max normalized and then fused with this area-consistency score:

$$s_i^{\mathrm{final}} = (1-\lambda)\tilde{s}_i^{\mathrm{rerank}} + \lambda a_i, \tag{3}$$

where $\tilde{s}_i^{\mathrm{rerank}}$ is the normalized Reranker score and λ controls the contribution of the geometric compatibility prior; we set $\lambda = 0.5$ in all experiments. The highest-scoring candidate is finally selected as the reference image.

3.4. Reference-Controlled Visual Style Transfer Based on Attention Distillation

During generation, LRRGen takes the target semantic mask, structured text prompt, and retrieved reference image with its paired semantic mask as inputs. Text cross-attention provides semantic guidance, while ControlNet constrains the target spatial layout (Zhang et al., 2023). The target and reference masks are resized to each selected attention layer and compared patch by patch to construct semantic matching weights: target–reference pairs receive a weight of one when their class labels agree and zero otherwise. These weights restrict reference guidance to corresponding semantic regions.

To transfer fine-grained reference appearance, we adopt the sampling guidance introduced in Attention Distillation (Zhou et al., 2025), as illustrated in Figure 5. After each DDIM denoising step, the current latent and a reference latent noised to the same timestep pass through the frozen U-Net to compute the attention distillation loss. The loss gradient is backpropagated to the current latent, which is refined with Adam updates before the next denoising step; all model parameters remain fixed.

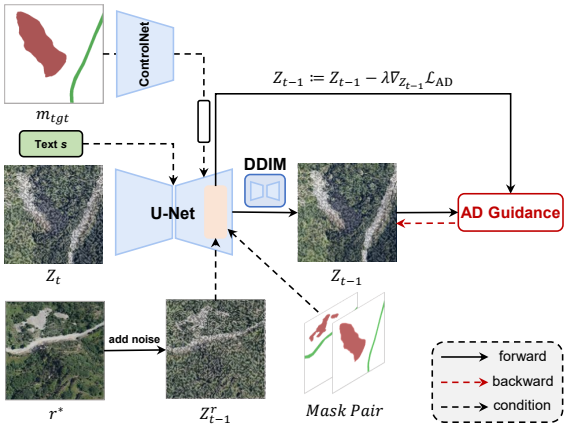


Figure 5: Diffusion sampling with guidance adopted from Attention Distillation (Zhou et al., 2025). Under text and ControlNet conditioning, the attention distillation loss guides latent updates using the retrieved reference and semantic masks.

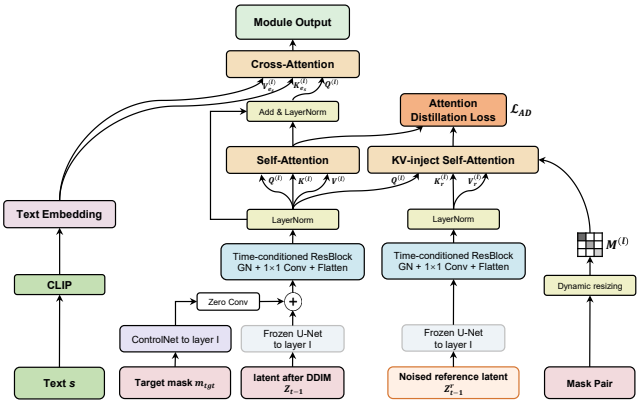


Figure 6: Mask-guided attention distillation loss. The current query and semantically matched reference keys and values form an attention target; its ℓ_1 discrepancy from the current self-attention output guides style transfer.

Figure 6 details the KV-based style transfer used in this adopted mechanism. At the last six self-attention layers, the current query Q is combined with reference keys K_r and values V_r, subject to the semantic matching weights, to form a reference-guided attention target. The attention distillation loss is the accumulated ℓ_1 discrepancy between the current self-attention outputs and these detached targets. This loss guides latent optimization to transfer reference textures within semantically matched regions while preserving the target layout.

4. Experiments

4.1. Dataset Construction

Data Sources. LRRGen uses the post-disaster UAV landslide image dataset constructed in this paper for training and evaluation. The data originates from UAV acquisitions of 13 typical landslide disaster areas in western China over the past 5 years. Due to variations in flight altitude, camera parameters, and imaging scales among the original images, we first uniformly resample all images to a 0.3 m Ground Sample Distance (GSD) and crop them into 512 × 512 pixel patches. Table 2 summarizes the acquisition locations, image dimensions, spatial resolutions, and image counts of the UAV data sources.

Table 2
Overview of UAV imagery data sources.

Location	Image Size (pixels)	GSD (m/pixel)	Original Images	Filtered Images
Aerzhai	5472 × 3648	0.10	75	74
Ganluo	5464 × 3956	0.12	1,237	169
Fa'er	5472 × 3648	0.10	319	82
Jinshajiang Baige	8688 × 5792	0.05	83	79
Jiuzhai Heyezhai	6000 × 4000	0.15	53	52
Jiuzhai Shangshizhai	6000 × 4000	0.15	50	50
Leshan	4864 × 3648	0.10	87	85
Lixian Shiziping	5280 × 3956	0.20	976	141
Luding Moxi	7952 × 5304	0.10	126	55
Maoxian Xinmo	6000 × 4000	0.15	49	48
Tianquan	5472 × 3648	0.12	91	47
Wenma	5464 × 3956	0.20	1,461	132
Ya'an	5464 × 3956	0.20	2,868	201

Data Filtering. Subsequently, based on DINOv2 feature similarity (Oquab et al., 2024), we remove redundant samples caused by flight strip overlap, and manually discard invalid samples such as pure vegetation, pure water surfaces, severely blurred images, and those lacking effective disaster information. Specifically, a cosine similarity threshold of 0.95 between DINOv2 feature embeddings is used to identify near-duplicate image patches for removal. After completing the annotation of landslide and road regions, small-scale samples with a landslide mask proportion of less than 1% are further removed. Following data cleaning, redundancy removal, semantic correction, and scale filtering, a final set of 8,421 valid image samples is obtained.

Semantic Mask Annotation. Each sample comprises a UAV image, a semantic mask, and a structured text description. Semantic masks for landslides and roads are interactively annotated by human annotators using Labelme (Wada, 2021), with semi-automatic assistance from the Segment Anything Model (SAM) (Kirillov et al., 2023), to provide spatial control conditions.

Textual Annotation. Each structured text condition consists of three components: Overall, Landslide, and Road, all generated using Gemini 3.1 Pro (Google, 2026). The Landslide and Road descriptions follow the Visual Style Taxonomy introduced in Section 3.2, with mask-guided visual prompts directing the model to describe the corresponding regions using the predefined attribute categories and controlled vocabularies. The Overall description summarizes the image as a whole, including viewing geometry, illumination, scene composition, and environmental context.

Training Augmentation. Online augmentation applies horizontal flipping, small random rotations, and minor random cropping jointly to images and semantic masks. The corresponding text component is removed when a class disappears after cropping or its area proportion falls below a threshold. This increases spatial diversity while preserving image–mask–text consistency without expanding the stored dataset.

Test Samples. We randomly select 100 samples from the 8,421 valid samples as the test set, leaving the remaining 8,321 samples as the training set. All compared methods are evaluated on these same 100 test samples, using identical text descriptions and semantic masks as conditional inputs to ensure experimental fairness.

Reference Database. The reference database introduced in Section 3.3 comprises 3,903 landslide images selected from the training set for their clear structures and relatively complete geomorphological textures. All reference images are feature-encoded via Qwen3-VL-Embedding-8B (Li et al., 2026) and stored as a vector index. This reference database provides fine-grained, realistic visual style guidance for image generation.

4.2. Implementation Details

Training Settings. Only the semantic-mask-conditioned ControlNet is fine-tuned; the pretrained Stable Diffusion v1.5 backbone (Rombach et al., 2022b), including the U-Net, VAE, and text encoder, remains frozen. Fine-tuning is performed at 512×512 resolution for 10,000 steps using AdamW, a learning rate of 10^{-5}, 400 warm-up steps, and cosine decay. Training uses bf16 precision on four RTX 4090 GPUs, with a batch size of 32 per GPU and two gradient-accumulation steps (effective batch size: 256). Training takes approximately 10.9 hours (44 GPU-hours), and the final checkpoint is used for evaluation.

Inference Settings. All methods use 100 DDIM steps (Song et al., 2021a) at 512×512 resolution with a classifier-free guidance scale of 7.5 (Ho and Salimans, 2021). LRRGen selects one reference from the Top-10 retrieved candidates through reranking and area-consistency fusion, and applies reference guidance at the 32×32 and 64×64 U-Net decoder attention layers. On a single RTX 3090, inference averages 120.95 s per image over the 100 test samples, including 15.59 s for retrieval and 105.36 s for generation, with peak GPU memory of 16.35 GiB. These timings exclude model loading and offline index construction.

Evaluation Metrics. We employ four types of metrics to evaluate the model performance. These metrics assess generation quality from three perspectives: image similarity, semantic preservation, and textual consistency. MMD, Precision, and Recall are computed in both SigLIP (Zhai et al., 2023) and DINOv2 (Oquab et al., 2024) feature spaces to assess semantic and visual distributional similarity.

- Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) measures the discrepancy between generated and real image distributions.
- Precision and Recall (Kynkäänniemi et al., 2019) measure generated-sample fidelity and coverage of real scene patterns, respectively.
- VQA-Score (Lin et al., 2024) assesses the semantic consistency, structural quality, and visual realism of generated images.
- Landslide-region IoU measures the spatial overlap between landslide regions segmented from generated images and the input semantic mask.

4.3. Comparative Evaluation

Compared methods. We compare LRRGen with baseline, remote sensing, and reference-guided methods. Baseline methods include Vanilla ControlNet (Zhang et al., 2023), LoRA + ControlNet (Hu et al., 2022), and SD Fine-tuned (+ ControlNet). Remote sensing methods include DiffusionSat (Khanna et al., 2024) and Text2Earth (Liu et al., 2025), both combined with ControlNet for spatial conditioning. The reference-guided method is IP-Adapter + ControlNet (Ye et al., 2023), which provides image-based appearance guidance.

LoRA + ControlNet and SD Fine-tuned (+ ControlNet) first adapt the diffusion backbone through LoRA and direct fine-tuning, respectively, and then train a ControlNet branch; all other methods train only ControlNet while keeping their original pretrained backbones frozen. Thus, all methods undergo task-specific fine-tuning, with consistent training data and settings for ControlNet across methods, while backbone adaptation differs.

Quantitative analysis. In terms of image similarity, Table 3 shows that LRRGen achieves the highest Precision and Recall and the lowest MMD in both SigLIP and DINOv2 feature spaces on the 100 test samples. Its MMD values are 0.73 and 0.29, respectively, indicating closer agreement with the real-image distribution than the baseline, remote sensing, and reference-guided methods. These results suggest that the retrieved visual priors improve both the fidelity and coverage of generated UAV landslide imagery.

For semantic layout preservation, LRRGen achieves the highest landslide-region mIoU of 0.65, compared with 0.62 for Vanilla ControlNet and 0.57 for IP-Adapter + ControlNet (Table 3). This indicates better agreement between the generated landslide regions and the input semantic masks under the fixed segmentation evaluator. Together with the distribution metrics, these results show that LRRGen improves visual similarity while maintaining the prescribed spatial layout.

For text-conditioned VQA evaluation, Table 4 shows that LRRGen obtains the highest Overall, Semantic, and Structural scores among the generation methods, reaching 6.23, 6.28, and 6.73, respectively. These scores indicate stronger agreement with the textual conditions and better

Table 3
Quantitative comparison of spatial consistency and generative distribution metrics for UAV landslide image generation.

Method	mIoU (landslide) ↑	SigLIP Prec ↑	SigLIP Rec ↑	SigLIP MMD ↓	DINOv2 Prec ↑	DINOv2 Rec ↑	DINOv2 MMD ↓
Vanilla ControlNet	0.62	0.46	0.23	0.76	0.86	0.66	0.39
LoRA + ControlNet	0.60	0.37	0.13	0.86	0.69	0.49	0.68
SD Fine-tuned (+ ControlNet)	0.54	0.18	0.06	1.24	0.66	0.21	0.95
DiffusionSat (+ ControlNet)	0.47	0.19	0.20	1.06	0.72	0.66	0.53
Text2Earth (+ ControlNet)	0.59	0.08	0.26	1.61	0.36	0.64	1.33
IP-Adapter + ControlNet	0.57	0.36	0.17	0.84	0.92	0.39	0.60
Ours (Full)	**0.65**	**0.54**	**0.43**	**0.73**	**0.94**	**0.80**	**0.29**

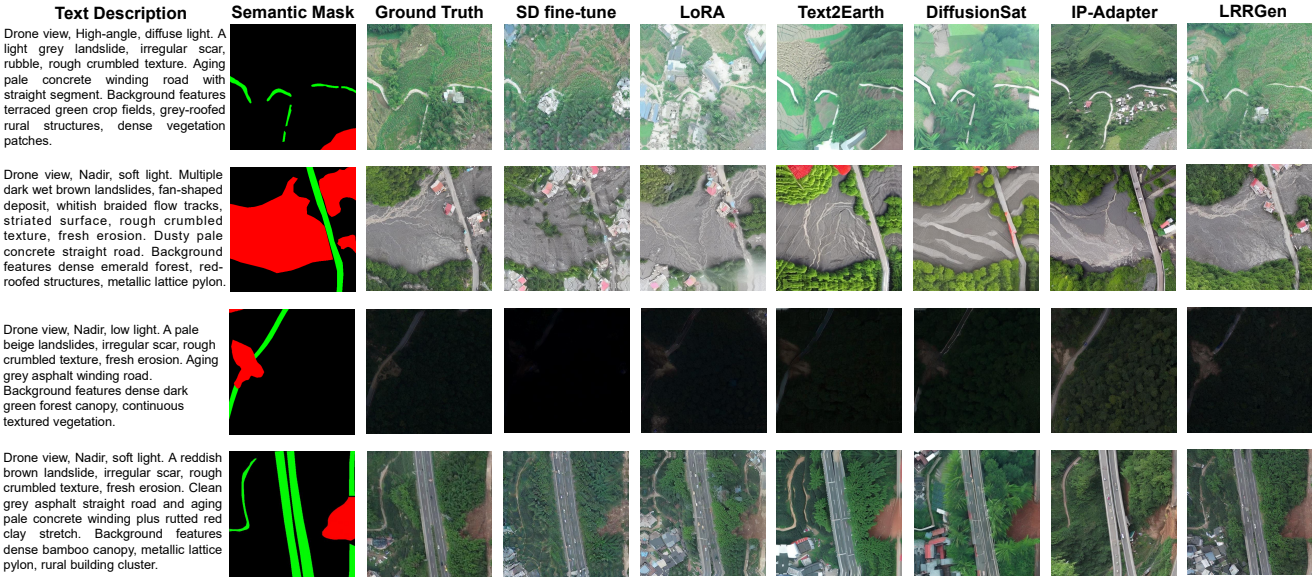


Figure 7
Qualitative comparison of different methods for text- and semantic-mask-conditioned post-disaster UAV landslide image generation.

Table 4
Detailed VQA-score comparison of different generation methods. VQA denotes the overall VQA-based quality score. Semantic, Structural, and Texture denote semantic consistency, structural completeness, and remote-sensing texture realism, respectively. GT (Real Data) is used only as a real-data reference baseline.

Method	VQA ↑	Semantic ↑	Structural ↑	Texture ↑
GT (Real Data, ref)	6.74	6.73	7.17	6.51
Vanilla ControlNet	5.71	5.73	6.34	5.13
LoRA + ControlNet	5.50	6.26	5.76	4.55
SD Fine-tuned (+ ControlNet)	4.28	4.84	4.38	3.68
DiffusionSat (+ ControlNet)	5.54	5.92	5.93	4.84
Text2Earth (+ ControlNet)	4.51	5.22	4.63	3.70
IP-Adapter + ControlNet	6.09	5.59	6.70	**6.15**
Ours (Full)	**6.23**	**6.28**	**6.73**	5.81

structural quality. IP-Adapter + ControlNet achieves a higher Texture score (6.15 versus 5.81), suggesting that LRRGen's advantage lies in its combined semantic and structural performance rather than texture quality alone.

Qualitative analysis. Figure 7 highlights the challenge of improving local texture realism while preserving the target scene layout. SD Fine-tuned (+ ControlNet) and LoRA + ControlNet produce coarse textures and fragmented boundaries, suggesting that backbone adaptation alone remains insufficient for this task. Text2Earth and DiffusionSat with ControlNet retain the broad layout but exhibit satellite-like colors and textures, revealing a gap between satellite imagery priors and UAV scene appearance. IP-Adapter + ControlNet adds visual detail, yet also shifts colors and textures across the scene, illustrating the limited regional selectivity of global reference guidance. LRRGen produces more realistic landslide textures while retaining road boundaries and surrounding vegetation, supporting the value of directing reference appearance to the corresponding semantic regions.

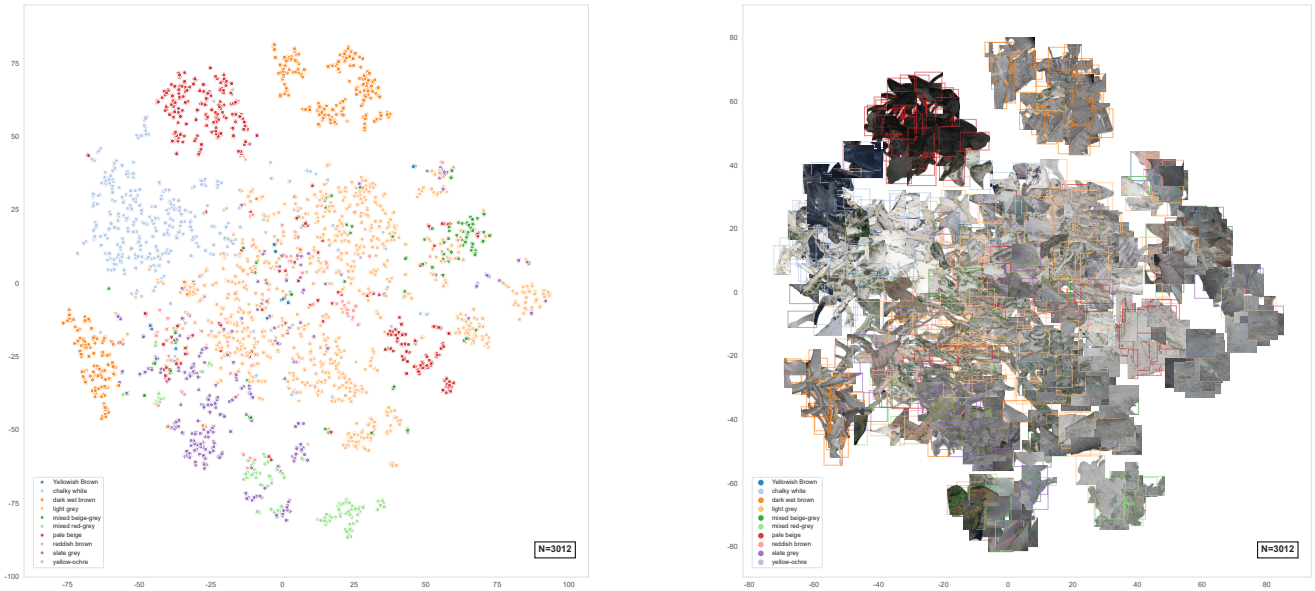
4.4. Ablation of the Retrival and Rerank Strategy

We evaluate reference selection while keeping the generation backbone, ControlNet, reference guidance, and sampling settings fixed. Table 5 compares two baselines with four Qwen-based variants. CLIP Matching uses general image–text embeddings, whereas StyleTag Matching uses structured style labels. The Qwen variants start from coarse retrieval and add area-ratio fusion, reranking, or both. Precision, Recall, and MMD measure the resulting image distributions rather than retrieval accuracy.

Qwen Coarse Retrieval yields lower MMD than both baselines in both feature spaces, suggesting that its selected

Table 5
Reference selection baselines and retrieval strategy ablations. The lower four rows compare Qwen Coarse Retrieval with variants adding Area Fusion, Rerank, or both. Bold indicates the best value in each column.

Method	SigLIP Prec ↑	SigLIP Rec ↑	SigLIP MMD ↓	DINOv2 Prec ↑	DINOv2 Rec ↑	DINOv2 MMD ↓
CLIP Matching	0.28	0.20	0.83	0.91	0.45	0.45
StyleTag Matching	0.34	0.27	0.86	0.83	0.65	0.68
Qwen Coarse Retrieval	0.46	0.34	0.76	0.93	0.56	0.33
+ Area Fusion	**0.55**	0.35	0.76	0.93	0.72	0.32
+ Rerank	0.52	**0.43**	0.73	0.93	0.74	0.31
+ Area Fusion and Rerank	0.54	**0.43**	**0.73**	**0.94**	**0.80**	**0.29**



(a) Color distribution of landslide patches **(b) Real patch mapping**

Figure 8: Visual style taxonomy analysis based on the embedding distribution of landslide patches.

references better match the target visual distribution. Adding Area Fusion improves SigLIP Precision and DINOv2 Recall, but slightly increases SigLIP MMD, indicating that compatible region proportions alone do not ensure appearance agreement. Reranking reduces MMD in both feature spaces and improves Recall, supporting the value of finer semantic and appearance matching.

Combining Area Fusion and Rerank achieves the lowest MMD in both feature spaces and the highest DINOv2 Precision and Recall, while matching the best SigLIP Recall. Although its SigLIP Precision is slightly below that of Area Fusion alone, the combined strategy provides the strongest overall distributional performance. This suggests that area compatibility and fine-grained reranking provide complementary cues for reference selection.

4.5. Analysis of Visual Style Control

4.5.1. Visual Style Discriminability

Limited discrimination using color alone. We analyze the embedding-space analysis of the landslide regions based on Qwen3-VL-Embedding-8B (Li et al., 2026). Semantic masks are used to extract landslide-region patches from the reference images, reducing background interference in the feature representations. Subsequently, Qwen3-VL-Embedding-8B is used to encode the visual features of the landslide patches, and t-SNE (van der Maaten and Hinton, 2008) is employed to map the high-dimensional embeddings onto a two-dimensional plane, which is used to observe the distribution structure of the controlled visual attributes within the feature space.

The legend in Figure 8(a) uses the landslide color categories defined by the Visual Style Taxonomy (Section 3.2), obtained by manually consolidating Gemini-generated descriptions. Each projected point is assigned a display color according to its annotated color category, so points with the same label share the same display color. As shown in Figure 8(a), samples within the same color category tend to cluster in the embedding space, indicating that color captures some differences in landslide visual style. However, obvious divergence still exists within the same color category, suggesting that a single color attribute is insufficient to completely describe the landslide appearance. As illustrated in Figure 8(b), adjacent samples are not only similar in color but also tend to share similar texture features.

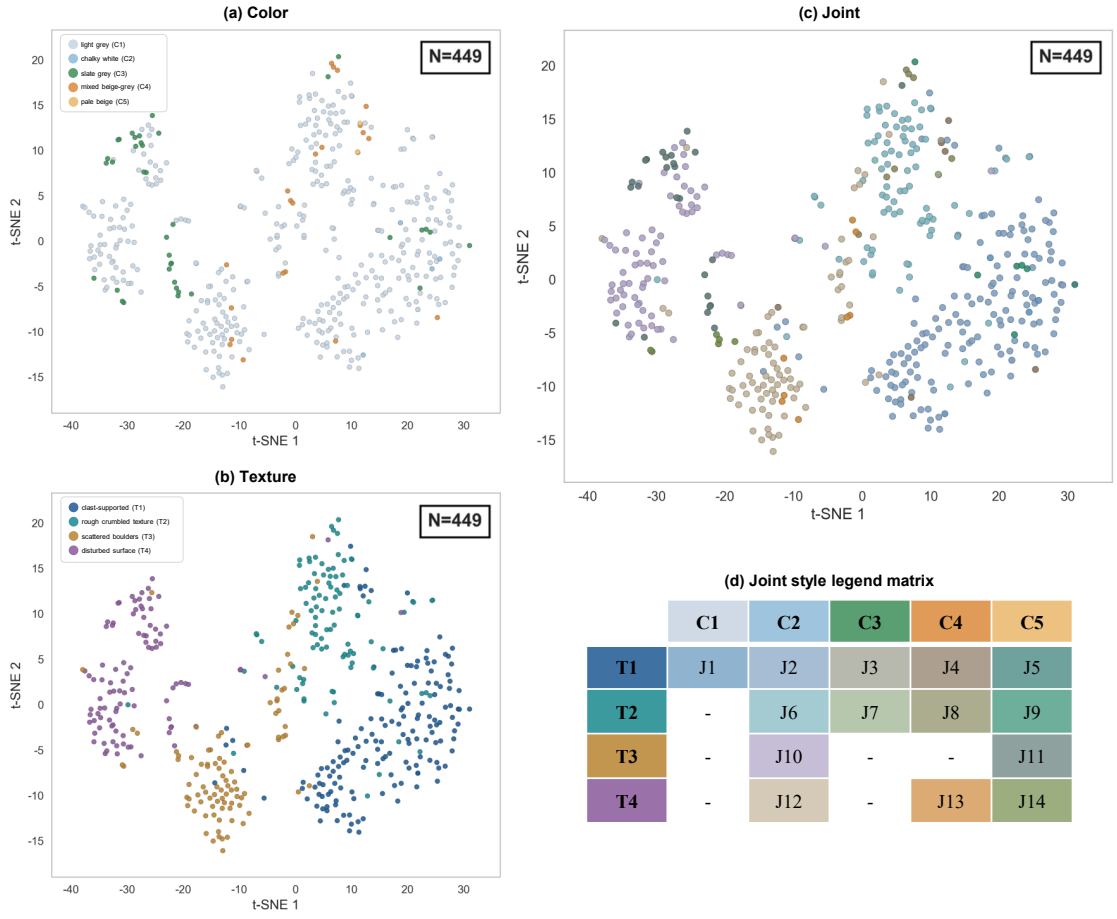


Figure 9: t-SNE visualization of landslide-region visual style embeddings under different attribute modeling schemes. (a) Color-based grouping. (b) Texture-based grouping. (c) Joint color–texture style grouping. (d) Category matrix of the joint visual styles. Each point represents a landslide-region patch encoded by Qwen3-VL-Embedding-8B, and its color indicates the corresponding attribute category or joint style category.

Improved discrimination through joint color and texture attributes. Selecting the Moxi Ancient Town region for local analysis, as shown in Figure 9, we compare the distributions of color attributes, texture attributes, and their joint representation. The results demonstrate that when relying solely on color attributes, samples still exhibit noticeable intra-class mixing; texture attributes provide additional discrimination of local surface patterns and material characteristics; and when color and texture attributes are jointly considered, the inter-class boundaries in the embedding space become clearer. These results indicate that color and texture attributes provide complementary visual information in the feature space, which also supports the use of a multi-attribute representation in constructing the proposed Visual Style Taxonomy.

4.5.2. Visual Style Controllability

Having established that the Visual Style Taxonomy distinguishes landslide appearances, we next examine whether its attributes can control the appearance of generated images. We fix the semantic mask and vary only the landslide color and texture descriptions, which guide reference retrieval and subsequent generation. Figure 10 shows that the generated landslide colors and textures change in accordance with the requested attributes, while the region's location and extent remain largely unchanged. This response supports the use of the taxonomy as a control interface: users can specify the desired landslide appearance through structured attributes while retaining the prescribed spatial layout.

4.6. Analysis of Semantic Control

Table 6 compares Vanilla ControlNet, IP-Adapter + ControlNet, and our method with and without semantic matching. The variant without SMatch removes only the semantic matching mask $\boldsymbol{M}$ in Figure 5, retaining reference-image KV injection (Zhou et al., 2025) and ControlNet conditioning to isolate its contribution to spatial control.

Although the variant without SMatch slightly improves image-based distribution metrics, its landslide IoU drops from 0.65 to 0.38. These gains come at the cost of unreliable semantic layout control, failing to meet our requirement for generation that follows the prescribed semantic mask. Semantic matching is therefore essential to preserving spatial

Figure 10: Controllable generation results of LRRGen under different landslide style attributes. Given a fixed semantic mask, LRRGen generates landslide regions with distinct colors, textures, and surface morphologies by modifying only the structured style description, while preserving the spatial position and extent of the landslide region.

Table 6
Quantitative ablation study of reference guidance methods

Method	Landslide IoU ↑	VQA Overall ↑	SigLIP Prec ↑	SigLIP Rec ↑	SigLIP MMD ↓	DINOv2 Prec ↑	DINOv2 Rec ↑	DINOv2 MMD ↓
Vanilla ControlNet	0.62	5.71	0.46	0.23	0.76	0.86	0.66	0.39
IP-Adapter + ControlNet	0.57	6.09	0.36	0.17	0.84	0.92	0.39	0.60
Ours w/o SMatch	0.38	6.34	0.58	0.44	0.72	0.96	0.84	0.28
Ours (Full)	0.65	6.23	0.54	0.43	0.73	0.94	0.80	0.29

constraints, even with a modest trade-off in image distribution metrics.

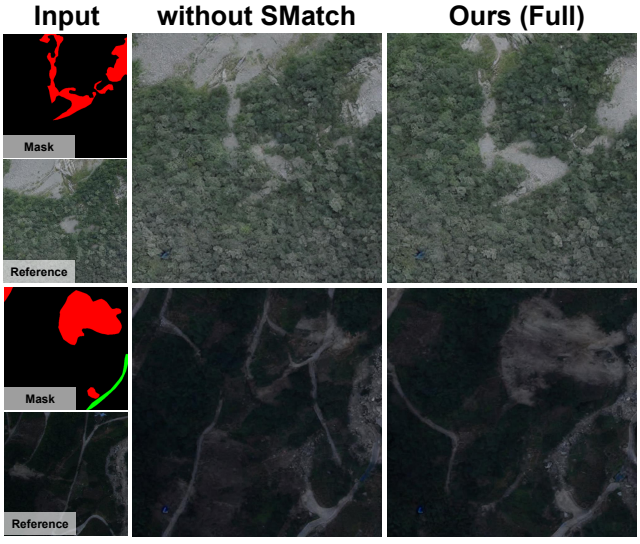


Figure 11: Comparison of semantic control with and without SMatch, focusing on alignment with the input semantic mask and reference texture transfer across semantic regions.

Enabling SMatch improves landslide IoU by 0.27, while SigLIP and DINOv2 MMD increase only from 0.72 to 0.73 and from 0.28 to 0.29, respectively (Table 6). The visual comparison in Figure 11 focuses on landslide-boundary alignment and cross-region texture leakage to complement these quantitative results.

4.7. Generalization Analysis on an Unseen Region

We evaluate LRRGen on a new landslide region excluded from training and the original test set. Of its 150 UAV images, 100 are added to the reference database and 50 serve as target cases. All model parameters remain frozen, testing adaptation through new references alone.

Figure 12 shows that LRRGen better reproduces regional landslide colors and textures than ControlNet, while preserving the target semantic layout more effectively than IP-Adapter + ControlNet. These results suggest that new reference images enable adaptation to unseen regional styles without additional training.

5. Conclusion

This paper proposed LRRGen, a retrieval-augmented controllable diffusion method for generating post-disaster UAV landslide imagery. The proposed method integrates structured visual style representation, Qwen-VL-based reference retrieval, ControlNet-based spatial constraints, and semantic-partitioned feature injection, thereby introducing local texture priors from real samples while preserving the target spatial layout. Experimental results demonstrate that LRRGen improves texture realism, distributional matching, and cross-region adaptability while maintaining spatial consistency, validating the effectiveness of structured style modeling, retrieval augmentation, and region-level style injection. Overall, LRRGen provides an effective solution for controllable synthesis and data augmentation of UAV-based disaster imagery.

Acknowledgements

This study was supported in part by the National Natural Science Foundation of China (Project No. U25A20772) and the Natural Science Foundation of Sichuan Province under Grant 2026NSFSCZY0054.

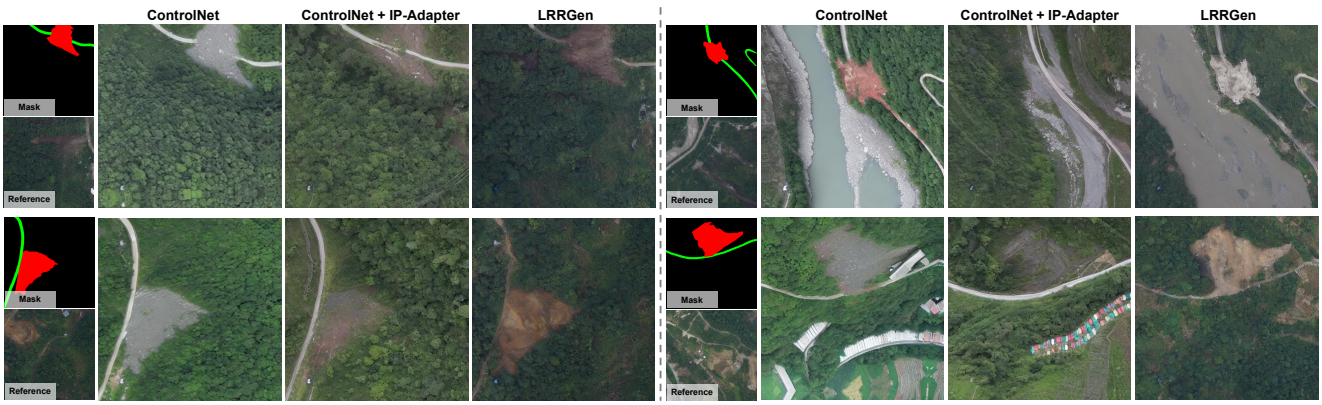


Figure 12: Qualitative comparison of unseen-region UAV landslide image generation. LRRGen better preserves the target semantic layout while producing local landslide textures and regional visual styles that are more consistent with the newly introduced reference images.

References

Chen, W., Hu, H., Saharia, C., Cohen, W.W., 2023. Re-imagen: Retrieval-augmented text-to-image generator, in: The Eleventh International Conference on Learning Representations.

Cheng, Z., Gong, W., Jaboyedoff, M., Chen, J., Derron, M.H., Zhao, F., 2025. Landslide identification in UAV images through recognition of landslide boundaries and ground surface cracks. Remote Sensing 17, 1900.

Chung, J., Hyun, S., Heo, J.P., 2024. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8795–8805.

Dong, R., Yuan, S., Feng, L., Zhang, J., Li, W., Chen, M., Luo, B., Zhang, W., Fu, H., 2025. Transferable image synthesis for remote sensing semantic segmentation via joint reference-semantic fusion. Information Fusion , 103839.

Duan, L., Zhao, S., Yan, W., Li, Y., Chen, Q.G., Xu, Z., Luo, W., Zhang, K., Gong, M., Xia, G.S., 2025. UNIC-Adapter: Unified image-instruction adapter with multi-modal transformer for image generation, in: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 7963–7973.

Gao, S., Xi, J., Li, Z., Ge, D., Guo, Z., Yu, J., Wu, Q., Zhao, Z., Xu, J., 2024. Optimal and multi-view strategic hybrid deep learning for old landslide detection in the loess plateau, northwest china. Remote Sensing 16, 1362.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in Neural Information Processing Systems, pp. 2672–2680.

Google, 2026. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-06-13.

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., 2012. A kernel two-sample test. Journal of Machine Learning Research 13, 723–773.

Gu, P., Jiang, Y., Lin, G., Hu, H., Jia, Y., Xiang, B., Liu, F., Chen, L., Ding, Y., Zhu, Q., 2026. Detect-Cluster-Reconstruct: Three-act on-board UAV landslide 3D modeling with multi-constraint clustering. IEEE Journal on Miniaturization for Air and Space Systems .

Guzzetti, F., Mondini, A.C., Cardinali, M., Fiorucci, F., Santangelo, M., Chang, K.T., 2012. Landslide inventory maps: New tools for an old problem. Earth-Science Reviews 112, 42–66.

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

Ho, J., Salimans, T., 2021. Classifier-free diffusion guidance, in: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations.

Hungr, O., Leroueil, S., Picarelli, L., 2014. The Varnes classification of landslide types, an update. Landslides 11, 167–194.

Khanna, S., Liu, P., Zhou, L., Meng, C., Rombach, R., Burke, M., Lobell, D., Ermon, S., 2024. Diffusionsat: A generative foundation model for satellite imagery, in: International Conference on Learning Representations, pp. 5586–5604.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R., 2023. Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026.

Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S., 2024. GeoChat: Grounded large vision-language model for remote sensing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27831–27840.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T., 2019. Improved precision and recall metric for assessing generative models, in: Advances in Neural Information Processing Systems.

Kyrkou, C., Theocharides, T., 2020. Emergencynet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13, 1687–1699.

Lee, C.T., Huang, C.C., Lee, J.F., Pan, K.L., Lin, M.L., Dong, J.J., 2008. Statistical approach to earthquake-induced landslide susceptibility. Engineering Geology 100, 43–58.

Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., Zhou, J., Lin, J., 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 .

Li, Z.H., Shi, A.C., Xiao, H.X., Niu, Z.H., Jiang, N., Li, H.B., Hu, Y.X., 2024. Robust landslide recognition using uav datasets: A case study in baihetan reservoir. Remote Sensing 16, 2558.

Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D., 2024. Evaluating text-to-visual generation with image-to-text generation, in: European Conference on Computer Vision, Springer. pp. 366–384.

Liu, C., Chen, K., Zhao, R., Zou, Z., Shi, Z., 2025. Text2earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. IEEE Geoscience and Remote Sensing Magazine .

Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J., 2024. RemoteCLIP: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing 62, 1–16.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of Machine Learning Research 9, 2579–2605.

Martha, T.R., Kerle, N., Jetten, V.G., van Westen, C.J., 2010. Characterising spectral, spatial and morphometric properties of landslides for semi-automatic detection using object-oriented methods. Geomorphology 116, 24–36.

MOT, 2010. Measures for reporting and handling information on transport emergencies. Ministry of Transport of the People's Republic of China, Jiao Yingji Fa [2010] No. 84. Article 3(8); in Chinese, English title translated by the authors.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M., 2022. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, in: Proceedings of the 39th International Conference on Machine Learning, PMLR. pp. 16784–16804.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2024. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research .

Pan, J., Lei, S., Fu, Y., Li, J., Liu, Y., Sun, Y., He, X., Peng, L., Huang, X., Zhao, B., 2025. Earthsynth: Generating informative earth observation with diffusion models. arXiv preprint arXiv:2505.12108 .

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., Wolf, T., 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR. pp. 8748–8763.

Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M., Murphy, R.R., 2021. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. IEEE Access 9, 89644–89654.

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

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022a. High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022b. Stable diffusion. https://github.com/CompVis/stable-diffusion. A latent text-to-image diffusion model.

Sastry, S., Khanal, S., Dhakal, A., Jacobs, N., 2024. Geosynth: Contextually-aware high-resolution satellite image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 460–470.

Sebaq, A., ElHelw, M., 2024. Rsdiff: Remote sensing image generation from text using diffusion model. Neural Computing and Applications 36, 23103–23111.

Shalev-Arkushin, R., Gal, R., Bermano, A.H., Fried, O., 2026. Imagerag: Dynamic image retrieval for reference-guided image generation, in: The Fourteenth International Conference on Learning Representations.

Shtedritski, A., Rupprecht, C., Vedaldi, A., 2023. What does clip know about a red circle? visual prompt engineering for vlms, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11987–11997.

Song, J., Meng, C., Ermon, S., 2021a. Denoising diffusion implicit models, in: International Conference on Learning Representations.

Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B., 2021b. Score-based generative modeling through stochastic differential equations, in: International Conference on Learning Representations.

Tang, D., Cao, X., Hou, X., Jiang, Z., Liu, J., Meng, D., 2024. Crs-diff: Controllable remote sensing image generation with diffusion model. IEEE Transactions on Geoscience and Remote Sensing 62, 1–14.

Taskina, L., Vorobyev, K., Abakumov, L., Kazarkin, T., 2026. BEHAVE-UAV: A behaviour-aware synthetic data pipeline for wildlife detection from UAV imagery. Drones 10, 29.

Toker, A., Eisenberger, M., Cremers, D., Leal-Taixé, L., 2024. Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 27695–27705.

Wada, K., 2021. Labelme: Image polygonal annotation with python.

Yang, L., Wang, Y., Li, X., Wang, X., Yang, J., 2023. Fine-grained visual prompting. Advances in Neural Information Processing Systems 36, 24993–25006.

Ye, H., Zhang, J., Liu, S., Han, X., Yang, W., 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 .

Yellapragada, S., Graikos, A., Triaridis, K., Prasanna, P., Gupta, R., Saltz, J., Samaras, D., 2025. Zoomldm: Latent diffusion model for multi-scale image generation, in: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 23453–23463.

Yu, Z., Liu, C., Liu, L., Shi, Z., Zou, Z., 2025. Metaearth: A generative foundation model for global-scale remote sensing image generation. IEEE Transactions on Pattern Analysis and Machine Intelligence 47, 1764–1781.

Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L., 2023. Sigmoid loss for language image pre-training, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11975–11986.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3836–3847.

Zhang, X., Huang, J., Zhang, L., 2026. Any2rsi: Controllable remote sensing text-to-image generation via any control and enriched description, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12852–12860.

Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K., 2023. Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in neural information processing systems 36, 11127–11150.

Zheng, Z., Tian, S., Ma, A., Zhang, L., Zhong, Y., 2023. Scalable multi-temporal remote sensing change data generation via simulating stochastic change process, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21818–21827.

Zhou, Y., Gao, X., Chen, Z., et al., 2025. Attention distillation: A unified approach to visual characteristics transfer, in: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 18270–18280.