RSCDG: Remote Sensing Change/Damage Image Generator Based on Prior Foundation Model and Multimodal Reference Information

Peng Chen [1], Guorui Ma [1], Haiming Zhang [1], Di Wang [1], Lunjun Fan [1]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China - pengchen0701@whu.edu.cn

Keywords: Remote Sensing Image Generation, Latent Diffusion Model, Multi-conditional Control, Change Detection, Damage Assessment.

Abstract

In scenarios such as natural disasters, military conflicts, and rapid urban expansion, high-quality post-event remote sensing images are often difficult to obtain in a timely manner, limiting the training and application of change detection, damage assessment, and related interpretation models. To address this issue, this paper proposes RSCDG, a remote sensing change/damage image generation framework based on prior foundation models and multimodal reference information. Built on a pretrained latent diffusion model, RSCDG integrates three types of conditional information: a Pre-event Visual Prompt Adapter extracts structural priors from the pre-event image via Prithvi-EO-2.0 to preserve background stability in unchanged regions; a Spatial Location Control Pathway introduces the change/damage mask into a ControlNet branch to improve spatial precision; and a Generation Content Text Controller uses a CLIP text encoder to guide semantically consistent generation. In addition, a Mask Alignment Loss is introduced to align the change patterns of generated and real images under the supervision of a frozen change detection model. Experiments on the LEVIR-MCI change scenario and the CEBD earthquake damage scenario show that RSCDG consistently outperforms ControlNet. In the change scenario, it achieves an FID of 28.92, an IS of 9.62, and a KID of 0.0139; in the damage scenario, the corresponding values are 37.82, 8.05, and 0.0187, respectively. Ablation results further confirm the effectiveness of the Mask Alignment Loss. Overall, RSCDG provides a practical solution for post-event sample construction, change detection data augmentation, and controllable sample generation for damage assessment.

1. Introduction

Remote sensing change detection (CD) and building damage assessment (BDA) are two core tasks in Earth observation and are of great value in applications such as natural disaster response, post-war assessment, and urban expansion monitoring. Early change detection methods mainly relied on traditional pipelines such as image differencing and ratio analysis, but they are often severely affected by complex textures, shadows, and pseudo-change noise in high-resolution remote sensing imagery. With the development of deep learning, remote sensing change detection has gradually shifted toward end-to-end learning frameworks. In particular, the emergence of fully convolutional Siamese architectures, Transformers, and long-range spatiotemporal modeling methods has substantially improved the representation of bitemporal differences (Bandara and Patel, 2022; Chen et al., 2022; Daudt et al., 2018; Fang et al., 2022; Zhang et al., 2020).

However, for both change detection and damage assessment, most discriminative models depend on sufficiently large, well-registered, and accurately annotated pre-event/post-event sample pairs. This prerequisite is often difficult to satisfy in real-world remote sensing applications. Post-event images are hard to acquire in time due to revisit cycles, cloud cover, and limited observation windows. Post-disaster and post-conflict scenarios are inherently sudden and scarce, resulting in high annotation costs. Moreover, high-quality bitemporal samples require precise labeling of changed regions, which is particularly critical in building damage assessment (Dong and Shan, 2013; Zhang et al., 2024; Gupta et al., 2019). Therefore, constructing high-quality, semantically credible, and spatially controllable post-event samples has become a key direction for alleviating the data bottleneck in remote sensing change-related tasks.

In recent years, diffusion models have gradually become a mainstream technical paradigm for high-fidelity image generation. DDPM establishes a mapping from noise to the data distribution through progressive noising and denoising (Ho et al., 2020), while LDM transfers diffusion modeling into latent space, reducing computational cost while preserving generation quality (Rombach et al., 2022). DiffusionSat demonstrates that diffusion models can serve as generative foundation models for satellite imagery (Khanna et al., 2024). CRS-Diff proposes a multi-condition framework for remote sensing image generation (Tang et al., 2024), and B3-CDG introduces diffusion generation into the construction of bitemporal building change samples (Chen et al., 2024). Although these studies provide an important basis for remote sensing generation, when the task becomes generating specified semantic change/damage content at specified locations given a pre-event image and a change mask, existing methods still face three insufficiently addressed challenges. First, background preservation remains inadequate, and the generator may introduce unnecessary texture re-rendering in unchanged regions. Second, spatial control and semantic control are prone to imbalance, since relying only on masks or text makes it difficult to jointly guarantee both where to change and what to change into. Third, explicit downstream task-oriented generation constraints are lacking. Conventional diffusion training may produce visually plausible results, but these are not necessarily well aligned with the task semantics required for change detection or damage recognition.

To address these issues, this paper proposes RSCDG, a multimodal conditional diffusion framework for remote sensing post-event change/damage image generation. The framework is designed around three objectives: background preservation, spatial localization, and semantic generation. Specifically, the remote sensing foundation model Prithvi-EO-2.0 is used to extract structural priors from the pre-event image, which are then

injected into the diffusion network through a Pre-event Visual Prompt Adapter. The change/damage mask is fed into a ControlNet (Zhang et al., 2023) branch to enable explicit spatial control over local regions. CLIP-based (Radford et al., 2021) text semantics are further introduced to specify the change type and damage state within the target region. In addition, a frozen BIT model is used to construct a Mask Alignment Loss, which enforces stronger alignment between the generated results and the real post-event images at the task level of change detection.

The main contributions of this paper are summarized as follows:

1. A multimodal conditional diffusion framework is proposed for remote sensing change/damage generation, which jointly models pre-event visual references, spatial masks, and textual semantics under a unified latent diffusion backbone.

2. A pre-event visual prompt adapter and a spatial control pathway are designed. The former leverages Prithvi-EO-2.0 to extract scene structural priors, while the latter applies ControlNet to impose explicit geometric constraints on changed regions.

3. A Mask Alignment Loss is introduced, where a frozen change detection model is used as an external task evaluator to explicitly align the generation process with bitemporal change semantics.

4. The effectiveness of the proposed method is validated on both change and damage scenarios. Experiments on the LEVIR-MCI and CEBD datasets show that the proposed method consistently outperforms the strong baseline ControlNet in terms of FID, IS, and KID.

2. Related Work

2.1 Evolution of Remote Sensing Change Detection Methods

High-resolution remote sensing change detection has evolved from traditional image processing to deep learning-based modeling. Traditional methods are typically based on pipelines such as pixel differencing and image transformation, but they are sensitive to noise and complex scene variations. With the advent of deep learning, research has shifted toward bitemporal representation learning. (Daudt et al., 2018) proposed FC-EF and FC-Siam-conc, which pioneered the systematic use of fully convolutional Siamese architectures for change detection. (Zhang et al., 2020) showed that IFN improves boundary integrity through dual-stream feature extraction and deep supervision. (Fang et al., 2022) showed that SNUNet-CD enhances multi-scale information fusion via dense connections. (Chen et al., 2022) introduced BIT, a compact spatiotemporal modeling mechanism driven by semantic tokens. (Bandara and Patel., 2022) proposed ChangeFormer, which learns long-range dependencies through a Siamese Transformer encoder. These discriminative methods have substantially improved change detection accuracy, but they still generally rely on sufficient real bitemporal samples.

2.2 Building Damage Assessment and Bitemporal Modeling in Disaster Scenarios

Unlike general urban change detection, building damage assessment focuses more on structural destruction after sudden disasters. xBD is currently one of the most widely used disaster remote sensing datasets (Gupta et al., 2019). (Dong and Shan, 2013) systematically summarized the challenges of earthquake-induced building damage detection, including image scale variation, geometric distortion, and annotation difficulty. At the model level, (Zhang et al., 2024) proposed SDCINet, which builds a cross-task fusion network between building object detection and change segmentation. Damage assessment is not merely a change detection problem; it also involves object recognition and structural damage interpretation. Therefore, generative models for damage scenarios must jointly consider geometric location control, texture realism, and structural damage semantics.

2.3 Diffusion Models and Remote Sensing Image Generation

Diffusion models have become an important paradigm in image generation. DDPM provides a unified framework for progressive noising and denoising (Ho et al., 2020), while LDM improves the efficiency of high-resolution generation through perceptual compression (Rombach et al., 2022). In the remote sensing domain, DiffusionSat demonstrates the potential of diffusion models for satellite image generation and related tasks (Khanna et al., 2024). CRS-Diff emphasizes controllable remote sensing generation under multiple conditions (Tang et al., 2024). B3-CDG combines latent diffusion with change masks and text prompts for pseudo-sample generation in bitemporal building change detection (Chen et al., 2024). Most existing studies focus on a single scenario and still lack a unified framework that simultaneously supports both change generation and damage generation while emphasizing background consistency and task-oriented constraints.

2.4 Multimodal Conditional Control and Remote Sensing Foundation Model Priors

Recent studies on controllable generation provide important motivation for this work. RemoteCLIP shows that remote sensing imagery can be aligned with textual semantics through large-scale vision-language pretraining, enabling meaningful remote sensing vision-language representations (Liu et al., 2024). GLIGEN shows that pretrained text-to-image diffusion models can be extended with grounding inputs for explicit control over the placement and identity of generated content (Li et al., 2023). T2I-Adapter further demonstrates that lightweight adapters can inject structural conditions into frozen diffusion backbones for efficient spatially guided generation (Mou et al., 2024). In remote sensing, SatMAE learns transferable temporal and multispectral representations through masked autoencoding (Cong et al., 2022), while CROMA shows the effectiveness of combining contrastive and reconstruction learning for multimodal remote sensing representation learning (Fuller et al., 2023).

3. Method Implementation

3.1 Overall Framework

The overall architecture of RSCDG is illustrated in Figure 1. Its generative backbone is a Stable Diffusion-style latent diffusion model. The real post-event image is first encoded into latent space via a VAE, followed by forward noising and reverse denoising. Meanwhile, the pre-event image x_{pre} is encoded by Prithvi-EO-2.0 (Szwarcman et al., 2024) and injected into the cross-attention layers of the diffusion UNet through the Pre-event Visual Prompt Adapter. The mask m is fed into the Spatial Location Control Pathway and progressively modulates the backbone features through a ControlNet branch. The text t is mapped into semantic tokens by a CLIP Text Encoder and participates in the denoising process together with the other

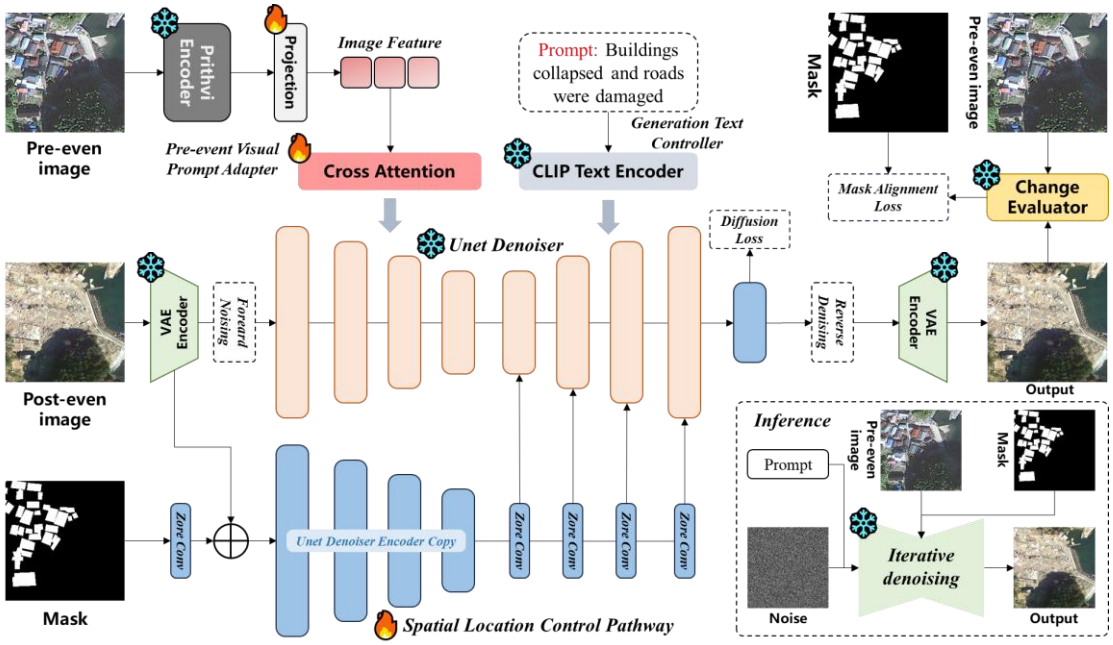

Figure 1. Overall framework of RSCDG. RSCDG is built upon a latent diffusion model, where the pre-event image, change mask, and text description are injected into the visual prompt branch, spatial control branch, and semantic control branch, respectively. During training, a frozen bitemporal change consistency evaluator imposes mask alignment supervision, enabling the generation of post-event remote sensing images with spatial controllability and semantic consistency.

conditions. During training, the model is jointly supervised by the diffusion reconstruction loss and the Mask Alignment Loss. During inference, it progressively generates the post-event image from random noise under multimodal conditional guidance.

From a design perspective, RSCDG explicitly decomposes functionality according to background preservation, spatial localization, and semantic generation. The visual reference pathway primarily addresses background stability, the spatial control pathway determines where changes occur, and the text pathway specifies the semantic attributes of the generated local content.

3.2 Latent Diffusion Modeling Under Multimodal Conditions

Within the latent diffusion framework, the real post-event image x_{post} is first encoded into a latent variable $z_0 = \mathcal{E}(x_{post})$ by the VAE. It is then progressively corrupted according to the diffusion timestep t to obtain z_t. Let the joint condition set be $y = (x_{pre}, m, t)$. The diffusion training objective is written as:

$$L_{LDM} = \mathbb{E} * (z * 0, y, \epsilon, t)[|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))|_2^2], \quad (1)$$

where $\epsilon \sim \mathcal{N}(0, I)$ denotes Gaussian noise, $\epsilon_\theta(\cdot)$ is the noise prediction network, and $\tau_\theta(y)$ denotes the multimodal conditional embedding function. Unlike standard text-to-image diffusion models, the conditions in this work form a composite control signal composed of the pre-event visual prompt, the spatial mask condition, and the text semantic condition. The conditional embedding process can be conceptually expressed as:

$$\tau_\theta(y) = \text{Fuse}\big(\phi_v(x_{pre}), \phi_m(m), \phi_t(t)\big), \quad (2)$$

where $\phi_v(\cdot)$, $\phi_m(\cdot)$, and $\phi_t(\cdot)$ denote visual reference encoding, mask-based spatial encoding, and text semantic encoding, respectively. Structured fusion is achieved by routing different condition types into different network locations: the visual condition is injected into cross-attention layers, the mask condition is introduced through the ControlNet branch, and the text condition follows the standard text-conditioning pathway.

3.3 Pre-event Visual Prompt Adapter

A frozen Prithvi-EO-2.0 encoder is used to encode the pre-event image x_{pre} and obtain remote sensing scene features $f_{pre} = \Phi_{\text{Prithvi}}(x_{pre})$. Following the design of IP-Adapter (Ye et al., 2023), f_{pre} is projected into visual prompt tokens compatible with cross-attention and injected into the UNet cross-attention layers in a decoupled manner alongside the text tokens. Let the original cross-attention be defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (3)$$

where Q is derived from the UNet query features, and K and V come from the conditional embeddings. While preserving the textual condition K_{text} and V_{text}, the proposed method introduces a visual prompt branch K_{vis} and V_{vis}, and concatenates them for attention computation. The core role of this module is to constrain the generator to perform local editing based on the pre-event image as the background foundation. In change scenarios, it preserves the original urban textures and road topology; in damage scenarios, it maintains stable structures in undamaged regions.

3.4 Spatial Location Control Pathway

A ControlNet-based Spatial Location Control Pathway is introduced to incorporate the change/damage mask as an explicit spatial condition into the diffusion denoising process. Specifically, the encoder structure of the Stable Diffusion UNet is replicated to form a ControlNet branch, and zero-initialized convolutions are used to inject mask-encoded features into the

backbone layer by layer. Let the original UNet feature be F_{unet} and the ControlNet branch output be F_{ctrl}. The injection process is expressed as:

$$F_{out} = F_{unet} + \mathcal{Z}(F_{ctrl}), \quad (4)$$

where $\mathcal{Z}$ denotes the zero-convolution layer. At initialization, $\mathcal{Z}(F_{ctrl}) \approx 0$, allowing the conditional branch to start from near-zero influence and gradually learn effective control. This mask constraint enables the model to more clearly distinguish between regions that should change and regions that should remain unchanged, thereby improving the geometric plausibility and boundary controllability of local generation.

3.5 Generation Text Controller

A frozen CLIP Text Encoder is used to encode the natural language description t into a semantic embedding $e_t = \Phi_{\text{CLIP}}(t)$, which is then fed into the UNet cross-attention module as the standard text condition. Pretrained on large-scale image-text alignment tasks, CLIP is able to represent language concepts such as *newly added*, *expanded*, *collapsed*, and *damaged* effectively (Radford et al., 2021).

3.6 Mask Alignment Loss

When optimized only with the diffusion reconstruction loss, the model tends to favor making the generated image look like a realistic post-event image, but not necessarily ensuring that it is consistent with the real post-event image in the sense of change detection. To alleviate this issue, a Mask Alignment Loss is introduced, where a frozen bitemporal change detection model M (implemented based on BIT) serves as an external task evaluator for supervision. Let

$$y_{true} = M(x_{pre}, x_{post}), y_{pred} = M(x_{pre}, \hat{x}_{post}), \quad (5)$$

then the Mask Alignment Loss is defined as:

$$L_M = -\frac{1}{N \cdot C} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{true,i,c} \log (y_{pred,i,c}), \quad (6)$$

where N is the number of pixels and C is the number of classes (change/unchanged). L_M encourages BIT to produce as similar a change pattern as possible for the two input pairs, namely pre-event/real post-event and pre-event/generated post-event, thereby correcting the generator at the level of change semantics.

By combining the diffusion reconstruction loss and the task-alignment constraint, the overall loss function is written as:

$$L = L_{LDM} + \lambda L_M, \quad (7)$$

where $\lambda = 0.3$ is used to balance visual quality and task consistency.

4. 4 Experiments

4.1 Datasets and Data Splits

LEVIR-MCI change scenario dataset: LEVIR-MCI is a multi-level change interpretation dataset proposed by (Liu et al., 2024), containing 10,077 pairs of bitemporal remote sensing images. Each image pair has a size of 256×256 and a spatial resolution of 0.5 m, and is annotated with both a pixel-level change mask and five change-description texts. We re-split the dataset into training, validation, and test sets using a 7:1:2 ratio, resulting in 7,053 training pairs, 1,008 validation pairs, and 2,016 test pairs.

CEBD damage scenario dataset: CEBD is a disaster-related building data resource constructed by (Zhang et al., 2024), focusing on building damage/change scenarios. This damaged-building dataset contains 10,377 samples and 52,661 damaged instances, covering building damage conditions across multiple regions and multiple disaster sources. We also split it into training, validation, and test sets using a 7:1:2 ratio, resulting in 7,264 training samples, 1,038 validation samples, and 2,075 test samples. The CEBD texts are automatically generated by InternVL3 (Zhu et al., 2025) and then manually reviewed. The prompt template for InternVL3 is as follows:

You are a professional remote sensing disaster interpretation expert. Given a pre-event image, a post-event image, and a binary damage mask, write one concise English sentence describing the visible building damage inside the masked region.

Requirements:

1. Focus only on building-related change or damage visible from the images.
2. Compare the pre-event and post-event images together with the mask.
3. Mention the extent or pattern of damage only when it is visually evident, such as collapsed buildings.
4. Do not speculate about casualties, causes, rescue actions, or information that is not visually observable.
5. Keep the sentence objective, concise, and annotation-style.
6. Output one sentence only.

4.2 Implementation Details and Fair Comparison Settings

A unified training configuration is used throughout this paper: the input size is 256 × 256, the total number of training steps is 10,000, the batch size is 16, the optimizer is AdamW, and the learning rate is 6×10^{-5}. DDPM is used as the sampler, and 50 denoising sampling steps are used during inference. Prithvi-EO-2.0, the Stable Diffusion UNet backbone, the VAE, and the CLIP Text Encoder are all frozen, and only the visual Adapter and the ControlNet branch are trained. The original ControlNet is adopted as the main baseline, using the same diffusion backbone, resolution, training steps, and sampling settings.

4.3 Evaluation Metrics

FID, IS, and KID are used to evaluate generation quality.

FID (Fréchet Inception Distance) measures the distribution distance between generated and real samples in feature space, where lower values are better:

$$\text{FID} = |\mu_r - \mu_g|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}\right) \quad (8)$$

where (μ_r, Σ_r) and (μ_g, Σ_g) denote the mean and covariance of real and generated samples, respectively, in the feature space of the Inception network.

IS (Inception Score) measures the class discriminability and diversity of generated results, where higher values are better:

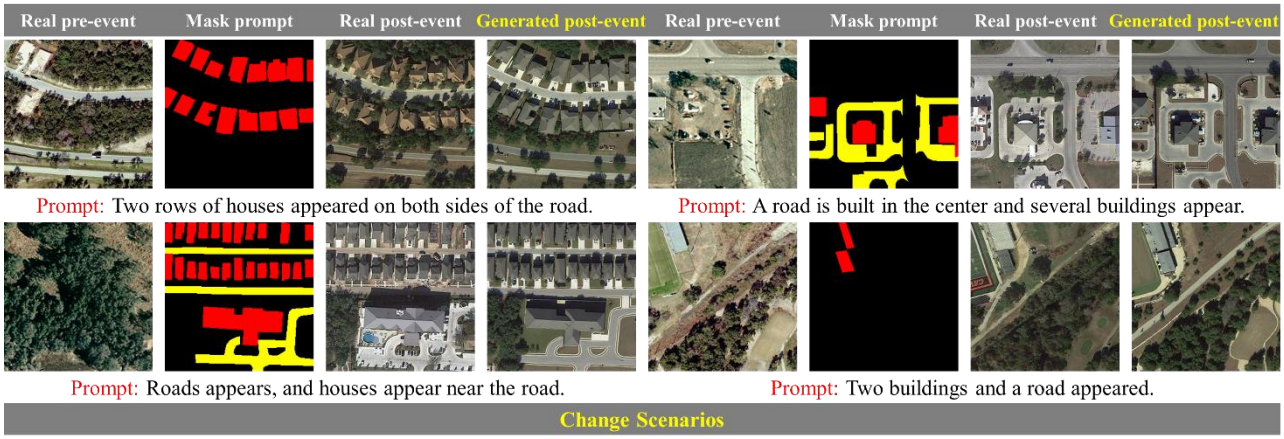

Figure 2. Visualization results in the change scenario on the LEVIR-MCI test set.

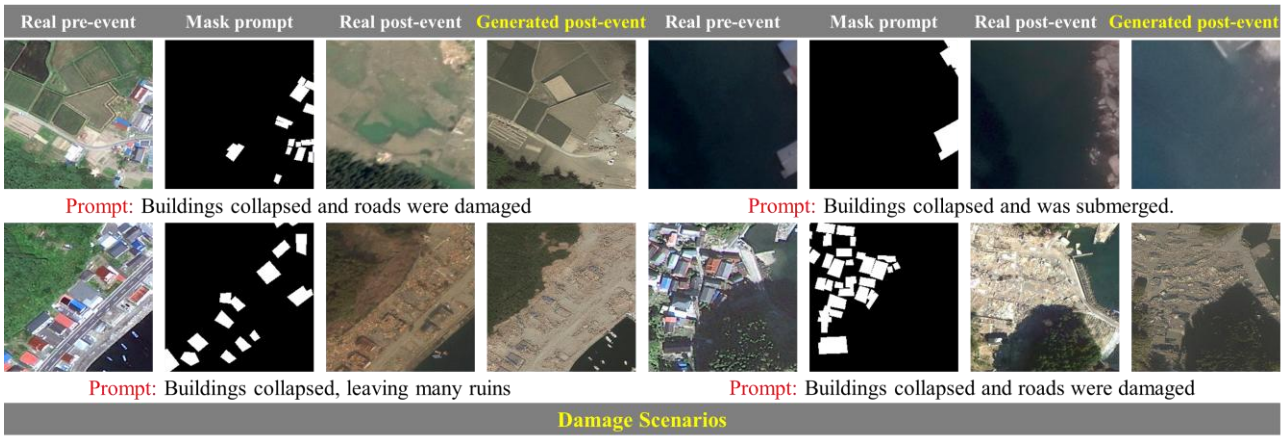

Figure 3. Visualization results in the damage scenario on the CEBD test set.

$$\mathrm{IS} = \exp\left(\mathbb{E}_{x\sim p_g}[\mathrm{KL}(p(y|x)|p(y))]\right), \tag{9}$$

where $p(y|x)$ is the class distribution predicted by the Inception network and $p(y)$ is the marginal distribution.

KID (Kernel Inception Distance) evaluates the discrepancy between real and generated feature distributions based on kernel methods, where lower values are better:

$$\mathrm{KID} = \widehat{\mathrm{MMD}}^2(p_r, p_g) = \mathbb{E} * x, x' \sim p_r[k(x,x')] + \mathbb{E} * y, y' \sim p_g[k(y,y')] - 2\mathbb{E}_{x\sim p_r, y\sim p_g}[k(x,y)], \tag{10}$$

where k is a polynomial kernel function.

4.4 Quantitative Comparison Results

4.4.1 Quantitative Results on the LEVIR-MCI Change Scenario: Table 1 reports the quantitative results on LEVIR-MCI. RSCDG clearly outperforms ControlNet on all three metrics: FID decreases from 57.16 to 28.92 (a 49.4% reduction), IS increases from 8.14 to 9.62 (an 18.2% improvement), and KID decreases from 0.0302 to 0.0139 (a 54.0% reduction). These results indicate that, in the change scenario, RSCDG explicitly preserves background structure through pre-event visual priors and focuses the generation capacity on the mask-constrained regions, thereby effectively suppressing the background drift commonly observed in the baseline model.

Method	FID (↓)	IS (↑)	KID (↓)
ControlNet	57.16	8.14	0.0302
RSCDG	28.92	9.62	0.0139

Table 1. Quantitative Results on the Change Scenario.

4.4.2 Quantitative Results on the CEBD Damage Scenario: Table 2 presents the results on the CEBD damage scenario. RSCDG reduces FID from 61.46 to 37.82 (a 38.5% reduction), improves IS from 7.89 to 8.05 (a 2.0% improvement), and lowers KID from 0.0353 to 0.0187 (a 47.0% reduction). The relatively limited gain in IS reflects the intrinsic complexity of the damage generation task, which involves structural breakage, texture degradation, and related irregular patterns. Nevertheless, the substantial improvements in FID and KID show that multimodal conditional collaboration and task-alignment supervision still effectively enhance generation quality.

Method	FID (↓)	IS (↑)	KID (↓)
ControlNet	61.46	7.89	0.0353
RSCDG	37.82	8.05	0.0187

Table 2. Quantitative Results on the Damage Scenario.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-33-2026

4.5 Visualization Analysis Results

4.5.1 Visualization Analysis in the Change Scenario: As can be seen from the Figure 2, the post-event images generated by RSCDG preserve the building layout, road topology, and texture characteristics of the pre-event images in unchanged regions, without exhibiting background drift. Within the mask-constrained regions, the roof textures, scales, and orientations of newly added buildings are consistent with the remote sensing characteristics in the real post-event images. The extended road segments also connect naturally to the surrounding road network in terms of direction, width, and surface texture. Overall, the generated change targets transition smoothly into the background environment.

4.5.2 Visualization Analysis in the Damage Scenario: In the damage scenario (Figure 3), RSCDG is able to generate clear destruction patterns such as building collapse and structural breakage within the mask region, while maintaining stable building outlines and background textures in undamaged regions. Across all four sample groups, the generated results do not incorrectly extend local damage to the entire image. The visual contrast between damaged and surrounding intact regions remains clear, and the complexity and irregularity of the damage textures are well represented. The overall visual plausibility is markedly better than that of the baseline method

4.6 Ablation Study

To verify the effect of the Mask Alignment Loss, L_M is removed in both scenarios, and the model is trained using only the diffusion reconstruction loss. The results are shown in Table 3. In the change scenario, removing L_M degrades FID from 28.92 to 33.12, reduces IS from 9.62 to 9.37, and increases KID from 0.0139 to 0.0173. In the damage scenario, FID degrades from 37.82 to 42.12, IS decreases from 8.05 to 7.87, and KID increases from 0.0187 to 0.0203.

Scenario	Change Scenario			Damage Scenario		
	FID (↓)	IS (↑)	KID (↓)	FID (↓)	IS (↑)	KID (↓)
RSCDG w/o L_M	33.12	9.37	0.0173	42.12	7.87	0.0203
RSCDG	28.92	9.62	0.0139	37.82	8.05	0.0187

Table 3. Ablation study of the Mask Alignment Loss

5. Discussion

The main application value of RSCDG lies in its role as a post-event sample generation engine that can provide controllable supplementary training data for change detection and damage assessment, especially in scenarios where real post-disaster samples are insufficient. Nevertheless, several limitations remain. First, RSCDG depends on reliable pre-event images and relatively usable mask conditions, making it more suitable for controllable generation under known target regions. In addition, the current training and validation are conducted at a fixed crop scale, and long-range consistency over larger spatial extents has not yet been fully verified. Second, the effectiveness of the Mask Alignment Loss is bounded by the capability of the external evaluator, and bias in the evaluation model may be propagated to the generator. Future work may proceed in three directions: incorporating larger-scale remote sensing foundation models to improve the adaptability of the visual reference pathway; forming a closed loop between the generator and the change detector to explore joint optimization of generation and detection; and investigating spatial autoregressive generation or cross-tile consistency constraints for large-scale remote sensing imagery, so that the model can better support more realistic region-level post-disaster analysis tasks.

6. Conclusions

This paper proposes RSCDG, a remote sensing change/damage image generation framework based on prior foundation models and multimodal reference information. Under a unified latent diffusion backbone, the framework jointly incorporates pre-event visual prompts, spatial location masks, and textual semantic conditions, and further constructs a Mask Alignment Loss using a frozen BIT change detection model, enabling the generation process to simultaneously account for visual realism, spatial controllability, and task-semantic consistency. Experiments on two representative scenarios, LEVIR-MCI and CEBD, show that RSCDG consistently outperforms the strong baseline ControlNet in terms of FID, IS, and KID, while the ablation study further verifies the effectiveness of the task-alignment loss. This work suggests that, for remote sensing post-event generation, the key lies in structurally organizing scene references, spatial constraints, semantic descriptions, and task supervision. RSCDG provides a practically meaningful technical path for data augmentation and controllable sample construction in change detection and damage assessment.

References

Bandara, W.G.C., Patel, V.M., 2022. A transformer-based Siamese network for change detection. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

Chen, H., Qi, Z., Shi, Z., 2022. Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing 60, 5604814.

Chen, P., Li, P., Wang, B., et al., 2024. B3-CDG: a pseudo-sample diffusion generator for bi-temporal building binary change detection. ISPRS Journal of Photogrammetry and Remote Sensing 218, 408-429.

Cong, Y., Khanna, S., Meng, C., et al., 2022. SatMAE: pre-training transformers for temporal and multi-spectral satellite imagery. In: Advances in Neural Information Processing Systems 35.

Daudt, R.C., Le Saux, B., Boulch, A., 2018. Fully convolutional Siamese networks for change detection. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 4063-4067.

Dong, L., Shan, J., 2013. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. ISPRS Journal of Photogrammetry and Remote Sensing 84, 85-99.

Fang, S., Li, K., Shao, J., et al., 2022. SNUNet-CD: a densely connected Siamese network for change detection of VHR images. IEEE Geoscience and Remote Sensing Letters 19, 1-5.

Fuller, A., Millard, K., Green, J.R., 2023. CROMA: remote sensing representations with contrastive radar-optical masked autoencoders. In: Advances in Neural Information Processing Systems 36.

Gupta, R., Hosfelt, R., Sajeev, S., et al., 2019. xBD: a dataset for assessing building damage from satellite imagery. arXiv preprint arXiv:1911.09296.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239.

Khanna, S., Liu, P., Zhou, L., et al., 2024. DiffusionSat: a generative foundation model for satellite imagery. In: International Conference on Learning Representations.

Li, Y., Liu, H., Wu, Q., et al., 2023. GLIGEN: open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22511-22521.

Liu, C., Chen, H., Qi, Z., et al., 2024. Change-Agent: towards interactive comprehensive remote sensing change interpretation and analysis from change detection and change captioning. arXiv preprint arXiv:2403.19646.

Liu, F., Chen, D., Guan, Z., et al., 2024. RemoteCLIP: a vision-language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing 62, 1-16.

Mou, C., Wang, X., Xie, L., et al., 2024. T2I-Adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. Proceedings of the AAAI Conference on Artificial Intelligence 38(5), 4296-4304.

Radford, A., Kim, J.W., Hallacy, C., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748-8763.

Rombach, R., Blattmann, A., Lorenz, D., et al., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684-10695.

Szwarcman, D., Roy, S., Fraccaro, P., et al., 2024. Prithvi-EO-2.0: a versatile multi-temporal foundation model for Earth observation applications. arXiv preprint arXiv:2412.02732.

Tang, D., Cao, X., Hou, X., et al., 2024. CRS-Diff: controllable remote sensing image generation with diffusion model. IEEE Transactions on Geoscience and Remote Sensing.

Ye, H., Zhang, J., Liu, S., et al., 2023. IP-Adapter: text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721.

Zhang, C., Yue, P., Tapete, D., et al., 2020. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing 166, 183-200.

Zhang, H., Ma, G., Fan, H., et al., 2024. SDCINet: a novel cross-task integration network for segmentation and detection of damaged/changed building targets with optical remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing 218, 422-446.

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.

Zhu, J., Wang, W., Chen, Z., et al., 2025. InternVL3: exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479.