This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ChangeBridge: Spatiotemporal Image Generation with Multimodal Controls for Remote Sensing

Zhenghui Zhao[1], Chen Wu[1†], Xiangyong Cao[2†],
Di Wang[3,4], Hongruixuan Chen[5], Datao Tang[2], Liangpei Zhang[1], Zhuo Zheng[6]
[1]State Key Laboratory of Information Engineering in Surveying,
Mapping and Remote Sensing, Wuhan University, China
[2]School of Computer Science and Technology, Xi'an Jiaotong University, China
[3]School of Computer Science, Wuhan University, China [4]Zhongguancun Academy, China
[5]Graduate School of Frontier Sciences, The University of Tokyo, Japan
[6]Department of Computer Science, Stanford University, USA

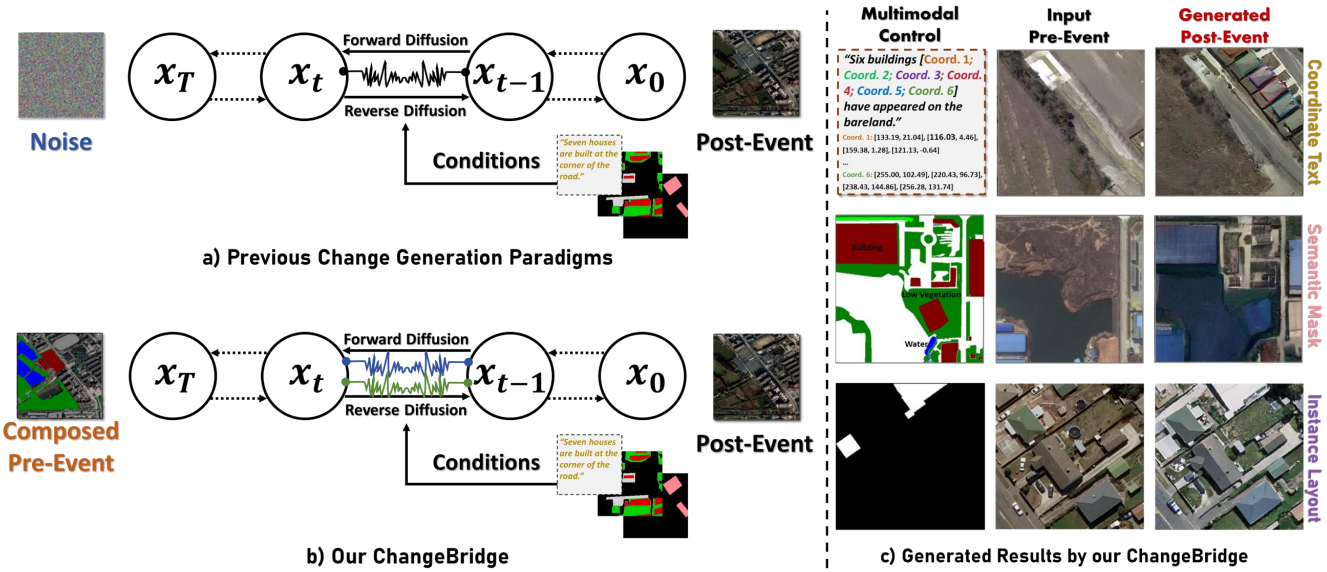


Figure 1. **Comparison of our ChangeBridge with previous change generation paradigms**. Unlike previous methods that rely on noise initialization, ChangeBridge starts from the composed pre-event state, with asynchronous drift diffusion, building a cross-spatiotemporal diffusion bridge process. It can generate post-event images based on given pre-event observations and multimodal controls, including **coordinate texts**, **semantic masks**, **or instance layouts**. Zooming in provides better visualization.

Abstract

Spatiotemporal image generation is a highly meaningful task, which can generate future scenes conditioned on given observations. However, existing change generation methods can only handle event-driven changes (e.g., new buildings) and fail to model cross-temporal variations (e.g., seasonal shifts). In this work, we propose ChangeBridge, a conditional spatiotemporal image generation model for remote sensing. Given pre-event images and multimodal event controls, ChangeBridge generates post-event scenes that are both spatially and temporally coherent. The core idea is a drift-asynchronous diffusion bridge. Specifically, it consists of three main modules: a) Composed Bridge Initialization, which replaces noise initialization. It starts the diffusion from a composed pre-event state, modeling a diffusion bridge process. b) Asynchronous Drift Diffusion, which uses a pixel-wise drift map, assigning different drift magnitudes to event and temporal evolution. This enables differentiated generation during the pre-to-post transition. c) Drift-Aware Denoising, which embeds the drift map into the denoising network, guiding drift-aware reconstruction. Experiments show that ChangeBridge can generate better cross-spatiotemporal aligned scenarios compared to state-of-the-art methods. Additionally, ChangeBridge shows great potential for land-use planning and as a data generation engine for a series of change detection tasks. Code is available at https://github.com/zhenghuizhao/ChangeBridge

†Corresponding author: Chen Wu and Xiangyong Cao

1. Introduction

Remote-sensing generation methods have been significantly advanced by the development of generative techniques. These advanced methods include layout-to-image synthesis [21, 37], modality transfer [2, 18], resolution modification [50, 56], and text-to-image generation [14, 36]. However, despite the diversity of these generative approaches, a key challenge is rarely explored: synthesizing future scenarios based on past observations and multimodal conditions, i.e., conditional spatiotemporal image generation. Addressing this challenge is crucial as it unlocks two critical capabilities. First, it provides a powerful "what-if" simulation tool for real-world applications like urban planning, land management, and scenario forecasting [19]. Second, it holds greater significance for the computer vision community, as it can function as a generative data engine. This can solve the severe data scarcity bottleneck for downstream spatiotemporal tasks (e.g., change detection), which require massive amounts of paired, pixel-aligned pre- and post-event data for training.

At its core, this spatiotemporal task is uniquely challenging because it must model a heterogeneous evolution. Specifically, the model must simultaneously generate: 1) drastic, event-driven changes in the foreground (e.g., new buildings emerging or regions being destroyed) and 2) subtle, temporal dynamics in the background (e.g., natural lighting shifts or vegetation growth).

Existing change generation models [51, 55, 56] follow the pipeline of Figure 1(a), which synthesizes changes by modifying spatial conditions, achieving the generation of event-driven changes. However, without cross-temporal modeling, they are unable to synthesize temporal dynamics. Therefore, these paradigms are ill-suited for handling this asynchronous, dual-natured generation.

To address this challenge, we propose ChangeBridge, a Drift-Asynchronous Diffusion Bridge designed to model this heterogeneous evolution. Instead of starting from pure noise, ChangeBridge directly bridges the pre-event and post-event states. Its architecture consists of three key modules: (a) composed bridge initialization, which starts the process from a composed pre-event state that coherently merges the pre-event background with the control-driven foreground; (b) asynchronous drift diffusion, which introduces a pixel-wise drift map to assign different evolution magnitudes to the foreground (high drift) and background (low drift); and (c) drift-aware denoising, where the denoising network is explicitly conditioned on this drift map to guide the differential reconstruction.

Functionally, ChangeBridge can generate realistic post-event images from a given pre-event image, allowing flexible multimodal event controls such as coordinate texts, instance layouts, or semantic masks. We evaluate two variants of ChangeBridge, utilizing three multimodal controls, on four datasets and compare their performance against six baselines. We demonstrate its effectiveness for future scenario simulation, and as a powerful data engine for downstream tasks, including change captioning and various forms of change detection.

Our contributions can be summarized as follows:

- We first introduce the task of conditional spatiotemporal image generation in remote sensing, simulating realistic future scenarios under multimodal controls.
- We propose ChangeBridge, a drift-asynchronous spatiotemporal diffusion bridge generative model, enabling conditions on instance layouts, semantic masks, and coordinate texts.
- We demonstrate that ChangeBridge can serve as a powerful generative data engine, significantly improving downstream change detection performance.

2. Related work

2.1. Spatiotemporal Scenario Simulation

Spatiotemporal scenario simulation plays a crucial role in understanding land use changes and urban dynamics in remote sensing. Traditional methods in this area primarily rely on rule-based algorithms [1, 7, 29] and statistical models [26, 39] to perform numerical simulations of spatiotemporal dynamics. Still, they fail to synthesize realistic future scenarios that are visually intuitive and easy to interpret.

Recently, generative models have revitalized scene-change generation. Several studies use GAN-based and diffusion-based frameworks to synthesize changes by modifying spatial conditions like maps, masks, or text-driven layouts [38, 51, 55, 56]. These methods focus on spatial differences, and cannot model cross-temporal transitions from historical observations. Lütjens et al. [19] made an initial attempt at spatiotemporal generation, using a GAN-based [43] framework for climate visualization, supporting the instance-layout condition. However, it can only simulate disaster weather from the current scene and falls short in terms of precise conditional control and universality. In this work, we first explore the general paradigm of spatiotemporal image generation with multimodal controls.

2.2. Multi-Conditional Generation

Multi-condition models, especially diffusion-based ones, are widely used for controllable generation. These models condition on spatial layouts [16], edges [54], depth maps [54], style exemplars [20, 32], and textual prompts [30]. They have also been extended to hierarchical layout control [6], multi-object image generation [41], and scientific structure design under complex property constraints [22].

Although these approaches achieve strong controllability, they still follow a noise-to-image diffusion process. During denoising, they usually rely on additional tech-

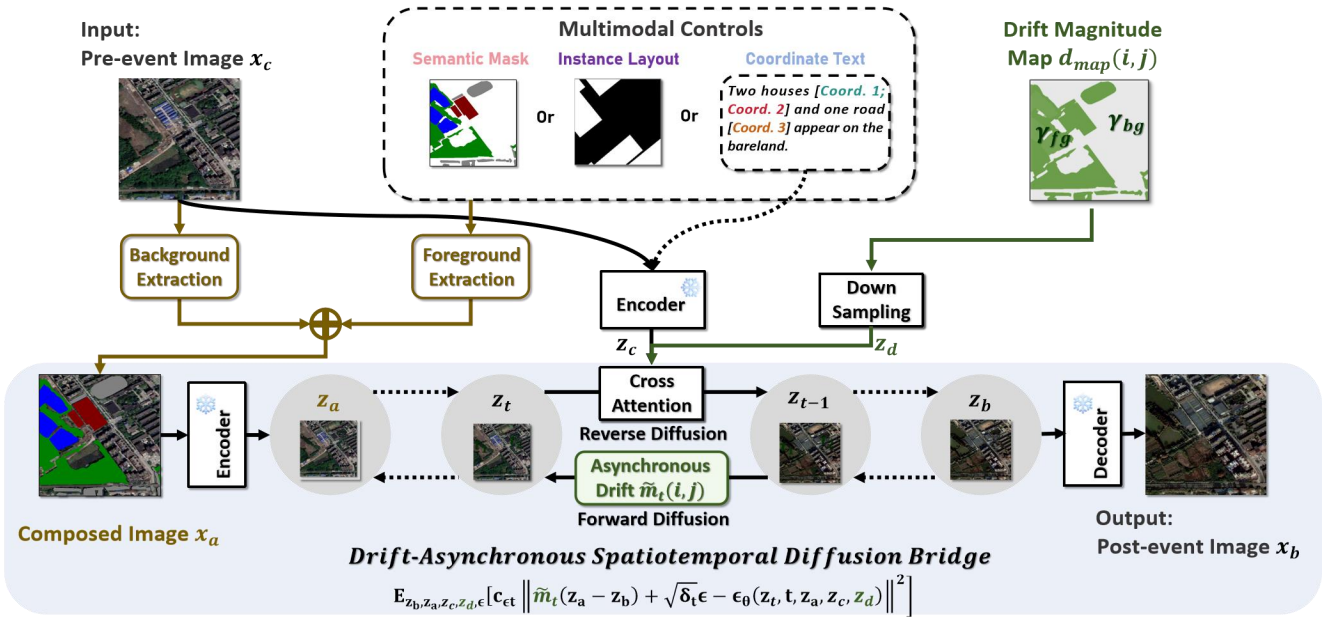


Figure 2. **Framework of ChangeBridge.** It includes three main modules: a) **Composed bridge initialization:** builds a cross-spatiotemporal diffusion bridge from a composed input, combining the pre-event background with the control-driven foreground. b) **Asynchronous drift diffusion:** introduces a pixel-wise drift map to assign different drift magnitudes to the foreground and background regions, enabling differentiated evolution. c) **Drift-aware denoising:** embeds the drift distribution into the denoising network to guide drift-aware reconstruction. Multimodal controls (e.g., coordinate text, semantic mask, and instance layout) are supported.

niques, such as classifier-free or guidance-free sampling [9], LoRA adaptation [47], or projected sampling [49], to align multiple conditions. Since generation starts from pure noise, the noise-to-image model must reconstruct both content and structure from scratch. It is difficult to maintain cross-image spatial coherence and temporal consistency. This limitation becomes especially evident in spatiotemporal generation, where transitions should remain consistent with historical observations under multiple controls.

2.3. Diffusion Bridge Model

The diffusion bridge formulation replaces the conventional "noise-to-state" generation paradigm with a "state-to-state" transformation, enabling explicit modeling of structured transitions between paired observations.

Recent studies have explored various bridge-based formulations, including Brownian Bridge diffusion models for image-to-image translation [15], denoising diffusion bridge models for stochastic distribution transport [57], and Schrödinger-bridge frameworks that generalize diffusion models for bidirectional distribution matching [8, 34]. Unlike conventional diffusion models that start from noise, bridge-based methods establish explicit mappings between two structured states or distributions, naturally preserving structural consistency and semantic alignment throughout the generation process [15, 34, 57]. Inspired by these advances and the need to preserve background structure in remote sensing, we propose a diffusion-bridge model for spatiotemporal generation with multimodal controls, enabling stable transitions between pre- and post-event images.

3. Methodology

We propose ChangeBridge, an asynchronous spatiotemporal diffusion bridge. ChangeBridge consists of three main components: a) Composed Bridge Initialization, as detailed in Section 3.2; b) Asynchronous Drift Modeling, as described in Section 3.3; and c) Drift-Aware Reverse Denoising, as presented in Section 3.4. The overall ChangeBridge framework is illustrated in Figure 2.

3.1. Preliminary

We first introduce the definition of diffusion bridge models, which extend the conventional diffusion generation from noise-based reconstruction to structured state transitions between two endpoints [15]. We adopt the classical Brownian bridge formulation, short for *diffusion bridge*.

Given two related images $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, a pre-trained autoencoder $\mathcal{E}$ encodes them into latent representations to reduce computational cost, i.e., $\boldsymbol{z}_a = \mathcal{E}(\boldsymbol{x}_a)$ and $\boldsymbol{z}_b = \mathcal{E}(\boldsymbol{x}_b)$, where $\boldsymbol{z}_a, \boldsymbol{z}_b \sim q(\boldsymbol{z})$. Following the diffusion bridge process [15, 25, 42], the latent state at timestep $t \in [0, T]$, evolving from $\boldsymbol{z}_0$ to $\boldsymbol{z}_T$, can be expressed as:

$$p(\boldsymbol{z}_t|\boldsymbol{z}_0, \boldsymbol{z}_T) = \mathcal{N}\left((1-\frac{t}{T})\boldsymbol{z}_0 + \frac{t}{T}\boldsymbol{z}_T,\ \frac{t(T-t)}{T}\boldsymbol{I}\right). \quad (1)$$

Forward process. The diffusion bridge constructs a stochastic trajectory that connects two structured endpoints $\boldsymbol{z}_b \rightarrow \boldsymbol{z}_a$ over T timesteps. Each intermediate latent $\boldsymbol{z}_t$ follows a Gaussian distribution:

$$q(\boldsymbol{z}_t \mid \boldsymbol{z}_b, \boldsymbol{z}_a) = \mathcal{N}(\boldsymbol{z}_t;\ (1-m_t)\boldsymbol{z}_b + m_t\boldsymbol{z}_a,\ \delta_t\boldsymbol{I}), \quad (2)$$

where $m_t = \frac{t}{T}$ denotes the drift term and $\delta_t = 2(m_t - m_t^2)$ represents the variance term along the bridge. This pro-

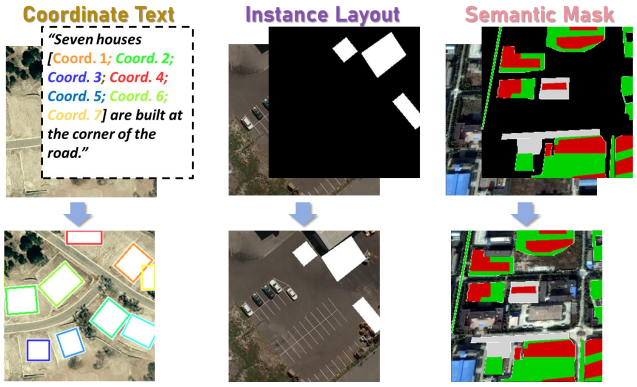


Figure 3. **Examples of composed images**. Each image combines the background from the pre-event image with the control-localized foreground.

cess defines a Markov chain. Unlike the standard diffusion model, which is a drift-free case of pure noise injection, the diffusion bridge introduces a non-zero drift m_t. This drift m_t progressively connects two structured endpoints, enabling a directional transition across timesteps.

Reverse process. The reverse process reconstructs the latent state $\boldsymbol{z}_b$ starting from $\boldsymbol{z}_a$ through a parameterized denoising network $\boldsymbol{\epsilon}_\theta$, modeled as:

$$p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t, \boldsymbol{z}_a) = \mathcal{N}\big(\boldsymbol{z}_{t-1};\, \hat{\boldsymbol{\mu}}_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_a),\, \hat{\sigma}^2_\theta(t)\boldsymbol{I}\big). \quad (3)$$

This reverse process learns the stochastic evolution from the latent state $\boldsymbol{z}_a$ to $\boldsymbol{z}_b$, directly modeling cross-state dependencies rather than learning from a random prior.

Training objective. The denoising network is trained to predict the perturbation $\boldsymbol{\epsilon}$ along the bridge trajectory:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_a, \boldsymbol{z}_b, \boldsymbol{\epsilon}}\left[\left\| m_t(\boldsymbol{z}_a - \boldsymbol{z}_b) + \sqrt{\delta_t}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_a)\right\|^2\right]. \quad (4)$$

Finally, the reconstructed latent $\boldsymbol{z}'_b$ is decoded by the pre-trained autodecoder $\mathcal{D}$ to generate the image $\boldsymbol{x}'_b = \mathcal{D}(\boldsymbol{z}'_b)$.

3.2. Composed Bridge Initialization

To enable spatiotemporal image generation, we construct composed images as the initialization of the diffusion bridge. This composition integrates the pre-event background and multimodal condition-guided foreground, providing context priors to jointly model foreground events and background temporal evolution. The composed image inputs are shown in Figure 3.

Conditional spatial awareness. Given a multimodal condition $\boldsymbol{x}_c$, which can be one of instance layouts, semantic masks, or coordinate texts. We localize event-relevant regions $\boldsymbol{M}_{\text{fg}} \in [0,1]^{1\times H\times W}$, and obtain a binary spatial prior according to the selected condition as follows:

- **Instance layout and semantic mask.** Each condition image maps event and no-event regions through color channels. We assume that a specific color value $\boldsymbol{c}_{\text{bg}}$ represents the no-event regions, i.e., background. The binary mask is then derived by pixel-wise comparison $\boldsymbol{M}_{\text{fg}}(i,j) = \begin{cases} 0, & \text{if } \boldsymbol{x}_c(i,j) = \boldsymbol{c}_{\text{bg}} \\ 1, & \text{otherwise} \end{cases}$. This definition ensures that all non-background event regions are explicitly highlighted as foreground.
- **Coordinate text.** Each coordinate-text pair, such as "Six buildings [Coords. 1–6] appear on the bareland," provides six spatially-localized coordinates of six buildings. For the k-th object in N_o coordinates, four corner points are given $\boldsymbol{P}_k = \{(x_1,y_1),(x_2,y_1),(x_1,y_2),(x_2,y_2)\}$, and $k = 1,\ldots,N_o$. Each $\boldsymbol{P}_k$ represents the Cartesian product of two horizontal and two vertical coordinates, which can form the four vertices of a rotated bounding box in the image plane. So we fit a rotated bounding box $\mathcal{B}_k = \text{RotBox}(\boldsymbol{P}_k) = \{(x,y) \mid \boldsymbol{R}(\theta_k)\boldsymbol{p} + \boldsymbol{t}_k \in \Omega_k\}$, where $\boldsymbol{R}(\theta_k)$ denotes a 2D rotation matrix with angle θ_k, $\boldsymbol{t}_k$ is the box center, and $\Omega_k \subset \mathbb{R}^2$ defines the local rectangular support. The foreground mask induced by all described objects is similarly defined as $\boldsymbol{M}_{\text{fg}}(i,j) = \begin{cases} 1, & \text{if } (i,j) \in \bigcup_k \mathcal{B}_k \\ 0, & \text{otherwise} \end{cases}$, This mapping $f_{\text{text}} : (\boldsymbol{P}_k, t_k) \mapsto \boldsymbol{M}^{\text{text}}_{\text{fg}}$ transforms coordinate-text pairs into a binary spatial mask, indicating event-relevant regions described by natural language.

Foreground–background extraction. According to the condition-aware foreground mask $\boldsymbol{M}_{\text{fg}} \in [0,1]^{1\times H\times W}$, we define the complementary background mask as $\boldsymbol{M}_{\text{bg}} = 1 - \boldsymbol{M}_{\text{fg}}$. The background regions are extracted from the pre-event image $\boldsymbol{x}_0$ via element-wise multiplication $\boldsymbol{x}_{\text{bg}} = \boldsymbol{M}_{\text{bg}} \odot \boldsymbol{x}_0$. The corresponding foreground regions $\boldsymbol{x}_{\text{fg}}$ are preserved from the condition $\boldsymbol{x}_c$ that $\boldsymbol{x}_{\text{fg}} = \boldsymbol{M}_{\text{fg}} \odot \boldsymbol{x}_c$, where $\odot$ denotes element-wise multiplication.

Composed image input. By combining the condition-driven foreground and pre-event background, we obtain the composed initial state of the diffusion bridge $\boldsymbol{x}_a = \boldsymbol{x}_{\text{fg}} + \boldsymbol{x}_{\text{bg}}$. This composition serves as the initial state of the diffusion bridge.

3.3. Asynchronous Drift Diffusion

Formally, let $\boldsymbol{x}_a$ denote the composed bridge input introduced above, obtained by merging the condition-driven foreground with the pre-event background. Let $\boldsymbol{x}_b$ denote the post-event remote sensing image. Their latent representations $\boldsymbol{z}_a = \mathcal{E}(\boldsymbol{x}_a)$ and $\boldsymbol{z}_b = \mathcal{E}(\boldsymbol{x}_b)$ are obtained through an autoencoder $\mathcal{E}$, where $\boldsymbol{z}_a, \boldsymbol{z}_b \in \mathbb{R}^{n\times m}$. We assume that the source and target latent variables follow two Gaussian distributions, $\boldsymbol{z}_a \sim q_{\text{pre}}(\boldsymbol{z}_a)$ and $\boldsymbol{z}_b \sim q_{\text{post}}(\boldsymbol{z}_b)$, both approximated by $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Similarly, the multimodal condition $\boldsymbol{x}_c$ is encoded into the same latent space through a domain-specific encoder τ_θ, i.e., $\boldsymbol{z}_c = \tau_\theta(\boldsymbol{x}_c) \in \mathbb{R}^{n\times m}$.

The drift-asynchronous diffusion process models the spatially varying drift between the composed bridge latent

$\boldsymbol{z}_a$ and the post-event target latent $\boldsymbol{z}_b$, which is specifically described in detail as follows.

Drift magnitude map. Given the binary foreground mask $\boldsymbol{M}_{\text{fg}} \in \{0,1\}^{H\times W}$ derived from multimodal conditions, we construct a pixel-wise drift magnitude map $\boldsymbol{d}_{\text{map}}$ to control the local evolution intensity during the diffusion process:

$$\boldsymbol{d}_{\text{map}}(i,j) = \boldsymbol{M}_{\text{fg}} \cdot \gamma^{fg} + \big(1 - \boldsymbol{M}_{\text{fg}}\big) \cdot \gamma^{bg}, \tag{5}$$

where γ^{fg} and γ^{bg} denote the drift magnitudes for the foreground and background evolutions, respectively, with $\gamma^{fg} > \gamma^{bg}$ by design. Then, the resulting map $\boldsymbol{d}_{\text{map}}$ is downsampled by adaptive average pooling to obtain the latent of the drift magnitude map $\boldsymbol{z}_d$, and is further normalized to $[0,1]$ for numerical stability. This drift map latent $\boldsymbol{z}_d$ is spatially aligned with the latent variables $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$, providing consistent modulation across regions.

Forward process. Building upon the diffusion bridge formulation in Eq. (2), we introduce a spatially asynchronous drift map $\boldsymbol{d}_{\text{map}}$ that modulates the drift magnitude across different spatial regions.

Specifically, for each pixel location (i,j), we redefine the drift coefficient as $\tilde{m}_t(i,j) = m_t \cdot \boldsymbol{z}_d(i,j)$, where $m_t = \frac{t}{T}$ is the canonical bridge coefficient defined in Eq. (2). This allows spatially adaptive drift velocities for region-specific spatiotemporal transitions. The forward transition can then be defined as:

$$q(\boldsymbol{z}_t \mid \boldsymbol{z}_b, \boldsymbol{z}_a) = \mathcal{N}\big(\boldsymbol{z}_t;\, (1 - \tilde{m}_t(i,j))\boldsymbol{z}_b + \tilde{m}_t(i,j)\boldsymbol{z}_a,\, \delta_t \boldsymbol{I}\big), \tag{6}$$

where $\delta_t = 2(m_t - m_t^2)$ is the variance term inherited from the original definition in Eq. (2). It is an empirically validated design to keep the variance schedule δ_t consistent with the original diffusion bridge for training stability. This design can be viewed as a drift-generalized version of the Brownian bridge diffusion [15], and the theoretically consistent counterpart with the asynchronous variance $\tilde{\delta}_t$ is provided in the supplementary.

3.4. Drift-Aware Denoising

During the reverse process, we further incorporate the spatially varying drift latent $\boldsymbol{z}_d$ to adaptively reconstruct the heterogeneous foreground and background dynamics.

Reverse process. Following the reverse dynamics in Eq. (3), the model reconstructs the target latent state $\boldsymbol{z}_b$ from the composed latent state $\boldsymbol{z}_a$ under the guidance of the spatially varying drift. The reverse transition is as follows:

$$p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t, \boldsymbol{z}_a, \boldsymbol{z}_c) = \mathcal{N}\big(\boldsymbol{z}_{t-1};\, \hat{\boldsymbol{\mu}}_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_a, \boldsymbol{z}_c, \boldsymbol{z}_d),\, \hat{\sigma}_\theta^2(t)\boldsymbol{I}\big), \tag{7}$$

where the drift term is modulated by the spatial magnitude map $\boldsymbol{z}_d$, allowing region-dependen, directional guidance during denoising. Here $\boldsymbol{z}_c$ denotes the pre-event image latent, providing the global context of the current observation scene. For the coordinate–text condition, we further concatenate the pre-event latent and the coordinate–text latent as $\boldsymbol{z}_c = [\,\boldsymbol{z}_{\text{text}} \,\|\, \boldsymbol{z}_{\text{pre}}\,]$, thereby integrating the pre-event context with event-relevant semantics.

Training objective. To learn the asynchronous dynamics, the denoising network $\boldsymbol{\epsilon}_\theta$ is trained to predict the perturbation noise $\boldsymbol{\epsilon}$ under the drift-modulated forward process. The loss is defined based on Eq. (4) with pixel-wise drift magnitude maps:

$$\mathcal{L}_{\text{asy}} = \mathbb{E}_{\boldsymbol{z}_a,\boldsymbol{z}_b,\boldsymbol{z}_c,\boldsymbol{\epsilon}}\Big[\big\|\tilde{m}_t(\boldsymbol{z}_a - \boldsymbol{z}_b) + \sqrt{\delta_t}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_a, \boldsymbol{z}_c, \boldsymbol{z}_d)\big\|^2\Big] \tag{8}$$

This objective modifies the drift term from Eq. 4, introducing a spatially adaptive drift coefficient while preserving the same variance schedule δ_t. As a result, the model learns heterogeneous transition behaviors, generating stronger evolution in event-related regions, and simulating smoother temporal dynamics in background areas.

3.5. Practical Implementation

Pipeline overview. This formulation is backbone-independent, making it applicable to both convolutional and transformer-based diffusion architectures. The pipeline follows as: 1) Given the *latent variable* $\boldsymbol{z}_t$, which evolves from the *composed image latent* $\boldsymbol{z}_a$, the model extracts multi-scale features and is guided by two additional signals during denoising: 2) the *pre-event latent* $\boldsymbol{z}_c$, providing global context information (for coordinate-text inputs, $\boldsymbol{z}_c$ additionally includes the text latent), and 3) the *drift latent* $\boldsymbol{z}_d$, indicating how strongly the foreground event and background temporal evolution occurs. 4) The *post-event latent* $\boldsymbol{z}_b$ is jointly guided by these signals to synthesize through the reverse denoising process.

Fusion mechanisms. We adopt standard condition-fusion strategies to integrate the drift latent $\boldsymbol{z}_d$ with the pre-event latent $\boldsymbol{z}_c$, depending on the network backbone.

For convolutional variants (e.g., UNet [28]), we perform conditional fusion via channel-wise concatenation of $\boldsymbol{z}_c$ and $\boldsymbol{z}_d$, forming a fused representation $\text{Concat}(\boldsymbol{z}_c, \boldsymbol{z}_d)$. The fused features are processed by convolutional layers, followed by a self-attention block to capture global context. The attention operation follows the standard formulation $\text{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\Big(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\Big)\boldsymbol{V}$. Here, $\boldsymbol{K}$ corresponds to the current latent feature, and $[\boldsymbol{Q}, \boldsymbol{V}]$ are derived from the fused representation. When coordinate–text conditions are available, cross-attention is further applied to align visual features with textual embeddings.

For transformer-based variants (e.g., DiT [23]), we follow its multimodal extension and adopt a FiLM-style additive modulation [24]. Specifically, the condition latents are concatenated and projected by a lightweight MLP $\phi(\cdot)$, and then added to the intermediate feature as a residual modulation: $\boldsymbol{x}_t' = \boldsymbol{x}_t + \phi(\text{Concat}(\boldsymbol{z}_c, \boldsymbol{z}_d))$. This design enables flexible condition injection without modifying the transformer architecture.

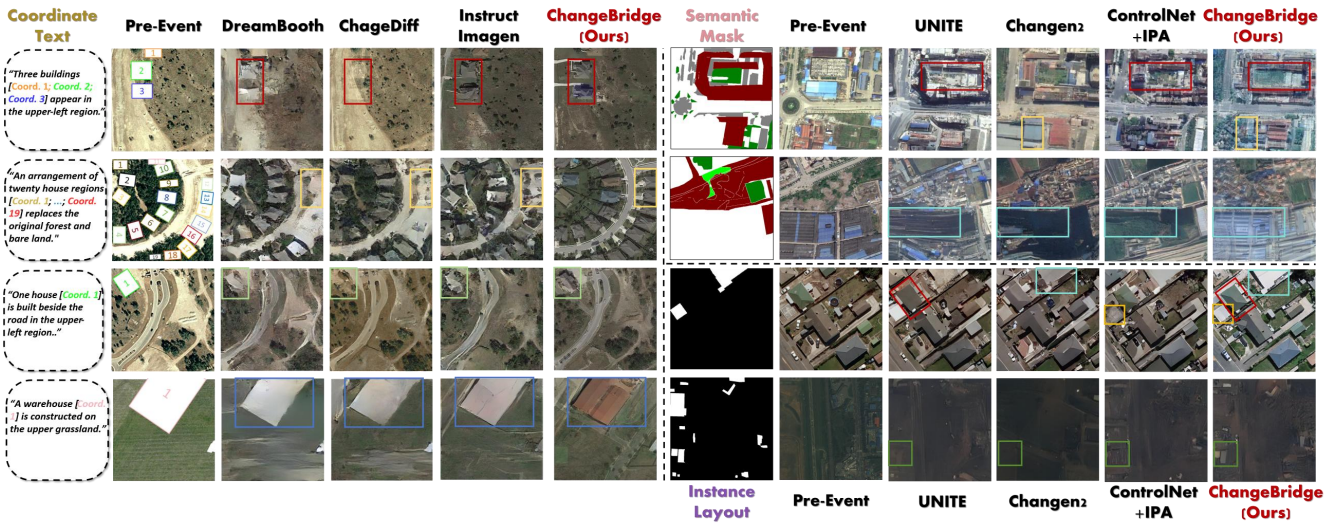


Figure 4. **Qualitative comparison** conditioned on **coordinate texts**, **semantic masks**, **and instance layouts**. Colored boxes highlight corresponding regions across methods (same color = same region; different colors = different regions). Zooming in provides better visualization.

4. Experiments

4.1. Experimental Setting

Dataset descriptions. We conduct experiments on four change detection benchmarks with default splits: 1) LEVIR-CC [3], for coordinate texts; 2) WHU-CD [13] and S2Looking [33], for instance layouts; and 3) SECOND [46], for semantic masks. We create a coordinate-augmented version of LEVIR-CC, adding object counting and precise object-level coordinates.

Implementation details. ChangeBridge uses both UNet- and DiT-based denoising backbones. The UNet variant is based on Stable Diffusion (SD) 1.5 [27], while the DiT variant uses DiT-XL/2 [23], initialized from pretrained weights and modulated with FiLM-style layers following SD3.5 [35]. Both variants are trained with the Brownian-Bridge objective, using 1,000 timesteps for training and 200 for inference. The UNet and DiT variants are trained for 60 and 100 epochs, taking approximately 12 and 25 hours, respectively, on 256×256 satellite images using Adam (1.0×10^{-4}) with a batch size of 64 on two NVIDIA A100 GPUs. The image encoder is a pretrained VQGAN [10], and the text encoder is SkyCLIP [44], fine-tuned for remote sensing. The maximum bridge variance is set to 1.0, optimized with L_1 loss. Asymmetric drift magnitudes are set to $\gamma^{fg} = 1.0$ and $\gamma^{bg} = 0.8$ for the WHU, S2Looking, and LEVIR-CC datasets, and $\gamma^{fg} = 1.0$ and $\gamma^{bg} = 0.7$ for SECOND. Unless specified, results use the DiT-based ChangeBridge, and ablations are conducted with the UNet variant.

4.2. Comparison with Prior Methods

Baseline models. We compare our ChangeBridge method against six existing methods: four multi-conditional generation methods (DreamBooth [30], Instruct-Imagen [12], UNITE [52], ControlNet [53] with IP-Adapter) [48], and two change generation methods (Changen2 [56] and ChangeDiff [51]). For multi-conditional methods, the pre-event image serves as the source, multimodal controls as the condition, and the post-event image as the target. For change generation methods, we use their original pipeline, inputting only pre-event images and controls.

Qualitative results. Figure 4 presents the qualitative comparisons of ChangeBridge, divided into three aspects: *1) Coordinate Text:* As shown on the left of Figure 4, DreamBooth often misaligns the generated buildings with the given coordinates, while Instruct-Imagen lacks fine-grained details. In contrast, ChangeBridge precisely follows the coordinate texts, preserves the scene structure, and produces clearer images. *2) Instance Layout:* As shown on the bottom-right of Figure 4, ChangeBridge aligns better with the provided spatial layouts than the other baseline methods. Compared to the change generation methods, it also produces more coherent cross-spatiotemporal evolution (e.g., lighting shifts) in the first example. *3) Semantic Mask:* As shown in the upper right of Figure 4, ChangeBridge demonstrates a clear improvement in semantic alignment over the three baseline methods across many regions. It also follows the instance layout constraints more accurately than the other three competing methods (Changen2, ControlNet+IPA, and UNITE), and preserves background consistency better.

Quantitative results. We quantitatively compare ChangeBridge as shown in Table 1. Following standard image-quality evaluation protocols, the performance is assessed using FID [11] and IS [31] across four datasets. For semantic consistency evaluation, we use SegFormer-based [45] mIoU/IoU for semantic-mask and instance-layout conditions, and CLIP-based cosine similarity (CosSim) for coordinate texts. The first row in each table reports the results of real data, as an upper bound. Note that FID scores in remote sensing are generally higher than those in natural-

Table 1. **Quantitative comparison** conditioned on instance layout, semantic mask, and text prompts. FID and IS are reported for all settings; mIoU (%) for semantic masks, IoU (%) for instance layouts, and cosine similarity (CosSim) for coordinate texts. Results on real data serve as the upper bound. ControlNet* denotes ControlNet+IPA, and Instruct* refers to Instruct-Imagen. Ours-U is our UNet variant, and Ours-T is the DiT variant.

Text	LEVIR-CC		
	FID ↓	IS ↑	CosSim ↑
Real Data	-	-	0.89
DreamBooth	95.64	1.52	0.76
ChangeDiff	55.60	3.43	0.79
Instruct*	48.17	3.70	0.81
Ours-U	**38.36**	**4.01**	**0.82**
Ours-T	**31.45**	**5.14**	**0.85**

Layout	WHU-CD			S2Looking		
	FID ↓	IS ↑	IoU ↑	FID ↓	IS ↑	IoU ↑
Real Data	-	-	81.30	-	-	82.47
UNITE	86.81	5.76	69.38	97.57	3.98	72.61
ControlNet*	52.08	5.12	71.54	94.68	4.56	75.20
Changen2	48.85	5.64	74.33	83.31	4.02	78.89
Ours-U	**45.47**	**5.88**	**75.30**	**72.56**	**4.60**	**78.45**
Ours-T	**40.12**	**6.77**	**78.13**	**56.42**	**5.22**	**79.40**

Semantic	SECOND		
	FID ↓	IS ↑	mIoU ↑
Real Data	-	-	76.19
UNITE	78.42	5.72	70.22
ControlNet*	90.81	3.31	72.30
Changen2	69.43	6.18	73.20
Ours-U	**62.24**	**6.03**	**73.47**
Ours-T	**59.33**	**6.41**	**74.26**



Figure 5. **Qualitative ablation** of ChangeBridge components. Starting from the SD1.5 baseline, we progressively incorporate CompBridge, AsyDrift, and DriftDenoise. Colored boxes highlight corresponding regions across methods (same color = same region; different colors = different regions).

Table 2. **Quantitative ablation** of ChangeBridge components. FID, IS, and IoU (%) are reported on WHU. CB denotes the CompBridge module, AD denotes AsyDrift, and DD denotes DriftDenoise.

CB	AD	DD	FID ↓	IS ↑	IoU ↑
			76.81	4.85	65.29
✓			56.24 -20.57	5.43 +0.58	71.87 +6.58
✓	✓		57.06 +0.82	5.61 +0.18	72.48 +0.61
✓	✓	✓	45.47 -11.59	5.88 +0.27	75.30 +2.82

image domains due to the feature distribution gap [55, 56].

ChangeBridge demonstrates outstanding performance with its variants across multiple conditional settings. *1) Coordinate Text:* It significantly outperforms Instruct-Imagen, achieving 31.45 (-16.72) FID, 5.14 (+1.44) IS, and 0.85 (+0.04) CosSim. These results highlight its superior cross-temporal generation quality and precise adherence to the provided coordinate texts. *2) Instance Layout:* When Evaluated on both WHU-CD and S2Looking datasets, ChangeBridge shows clear improvements. On S2Looking, it surpasses ControlNet+IPA with a 56.42 (-41.25) FID. Consistent with the visual results, ChangeBridge produces more realistic structures from pre-event inputs and matches the given spatial layouts more accurately. *3) Semantic Mask:* ChangeBridge outperforms the state-of-the-art Changen2, achieving 59.33 (-10.10) FID and 74.26% (+1.06%) mIoU, showing stronger coherence and class-consistency in evolution.

4.3. Ablation Study and Visualization Analysis

Effectiveness of components. We conduct ablation studies to validate the effectiveness of the main components of our model. As shown in Table 2, the introduction of *Composed Bridge Initialization* (CompBridge) significantly improves the SD1.5 baseline, reducing the FID by 20.57. AsyDrift continues to contribute to improvements in consistency. Additionally, it is notable that before the introduction of *Drift-Aware Denoising* (DriftDenoise), the FID slightly increased by 0.82, although it enhanced IS (+0.18) and IoU (+0.61%). This demonstrates the importance of embedding drift information in the reverse denoising process.

Visualizations also show significant improvements. As shown in Figure 5, the baseline method, SD1.5, shows limitations in both spatial control and spatiotemporal evolution (e.g., lighting transitions in the background). In contrast, our approach shows key advancements: 1) CompBridge aligns with the instance layout, generating a building with high accuracy. 2) AsyDrift mitigates cross-temporal inconsistencies, such as the disappearance of trees (highlighted in the yellow box). 3) Finally, DriftDenoise produces high-fidelity images, significantly aiding the reverse process and enhancing overall visual quality.

Visualization of inference evolution. We visualize the intermediate inference process of the drift-asynchronous diffusion bridge. This process shows how the model moves from current observations and multimodal controls to the

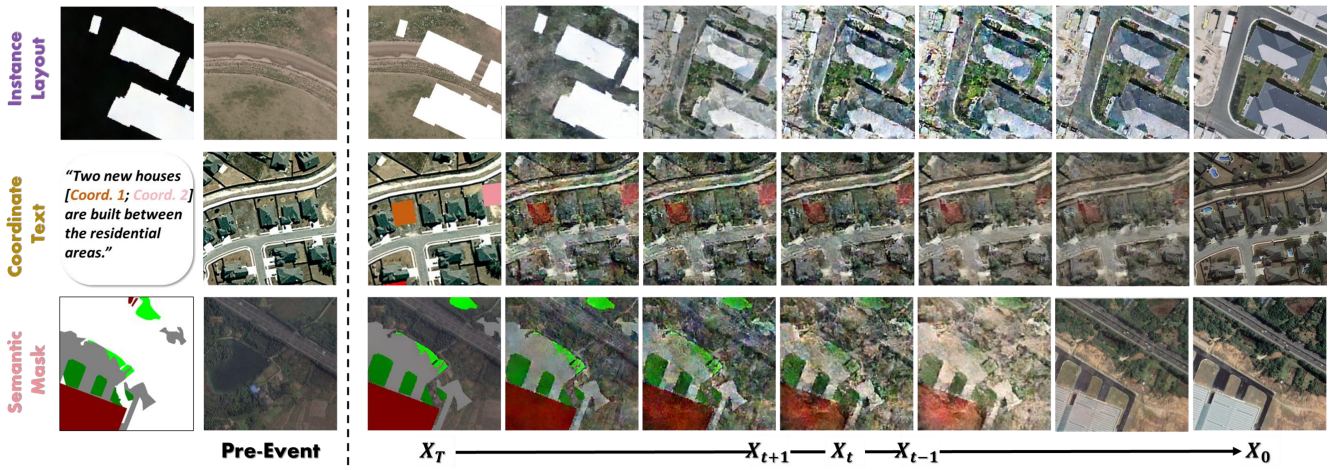


Figure 6. **Intermediate sampling process** of ChangeBridge. Our diffusion bridge starts from the composed pre-event state, and ends with the post-event state. Zooming in provides better visualization.

Table 3. **Performance on downstream change detection tasks**. BCD denotes binary change detection, SCD refers to semantic change detection, and CC is change captioning. IoU (%), mIoU (%), and CIDEr-D are reported.

Data	BCD task		SCD task	CC task
	WHU-CD	S2Looking	SECOND	LEVIR-CC
$\mathcal{B}$	72.39	46.94	73.33	134.12
$\mathcal{B} \cup \mathcal{B}'$	73.21$_{+0.82}$	**48.11**$_{+1.17}$	**74.02**$_{+0.69}$	138.50$_{+4.38}$
$\mathcal{B} \cup 2\mathcal{B}'$	**74.65**$_{+2.26}$	48.03$_{+1.09}$	73.57$_{+0.24}$	**145.09**$_{+10.97}$
$\mathcal{B} \cup 3\mathcal{B}'$	73.40$_{+1.01}$	47.45$_{+0.51}$	73.66$_{+0.33}$	142.31$_{+8.19}$

synthesized post-event images. As shown in Figure 6, the model preserves spatial and semantic consistency while evolving spatiotemporal dynamics. The foreground, like new buildings, evolves quickly and aligns with the layout, while the background changes more slowly, ensuring consistency with the pre-event image. This demonstrates that our asynchronous drift mechanism allows the foreground and background to evolve at different rates.

4.4. Evaluation on Change Detection Tasks

In this section, we evaluate the effectiveness of ChangeBridge as a data engine for three change detection tasks, using image augmentation during training. We apply different augmentation scales by varying the ratio between the original dataset $\mathcal{B}$ and the synthesized dataset $\mathcal{B}'$. For each task, we use the following synthesized data: instance-layout controls for binary change detection (BCD), semantic-mask samples for semantic change detection (SCD), and coordinate-text samples for change captioning (CC).

Binary change detection. Using BiT [4] on the WHU and S2Looking datasets, we evaluate with IoU. On WHU-CD, adding synthetic data improves IoU from 72.39% to 74.65% with a 2:1 synthetic-to-original ratio ($\mathcal{B} \cup 2\mathcal{B}'$), showing the benefits of synthetic data. On S2Looking, the optimal IoU of 48.11% is achieved with a 1:1 ratio ($\mathcal{B} \cup \mathcal{B}'$), resulting in a 1.17% gain over the baseline.

Semantic change detection. For this task, we use MambaSCD [5] on the SECOND dataset, evaluating with mIoU. Adding an equal amount of synthetic data ($\mathcal{B} \cup \mathcal{B}'$) improves mIoU slightly, from 73.33% to 74.02%, indicating improved semantic discrimination. However, further increases in the synthetic data ratio (beyond 1:1) lead to diminishing returns, with mIoU reaching 73.57% at the 2:1 ratio, suggesting potential overfitting.

Change captioning. Using RSICCformer [17] on the LEVIR-CC dataset and evaluating with CIDEr-D [40], the results show consistent improvement in CIDEr-D as synthetic data increases. The highest CIDEr-D score of 145.09 is achieved with a 2:1 ratio ($\mathcal{B} \cup 2\mathcal{B}'$), demonstrating that synthetic data enhances text-image alignment and improves caption generation. Overall, moderate augmentation scales (1:1 or 2:1 ratios) yield the best results across tasks, showing that synthetic data from ChangeBridge improves generalization in change detection.

5. Conclusion

We propose ChangeBridge, a conditional spatiotemporal generative model that generates realistic post-event scenarios from pre-event observations and multimodal controls. The model uses a drift-asynchronous spatiotemporal diffusion bridge. Experiments on four datasets and six baselines show that ChangeBridge achieves high-fidelity event synthesis. As a data engine, it also improves downstream change detection performance. This framework holds promising potential for applications in land-use planning and data-driven change analysis. In future work, we aim to extend ChangeBridge with a flow-matching formulation for more efficient and stable spatiotemporal generation.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (T2122014, 62272375, and 624B2109), the National Key Research and Development Program of China (2022YFB3903300), and the Key Technology Research Project of China National Petroleum Corporation (2025ZG82). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] V Chaturvedi and WT de Vries. Machine learning algorithms for urban land use planning: A review. *Urban Science*, 5(3):68, 2021. 2

[2] Bowen Chen, Liqin Liu, Chenyang Liu, Zhengxia Zou, and Zhenwei Shi. Spectral-cascaded diffusion model for remote sensing image spectral super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2

[3] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12 (10):1662, 2020. 6

[4] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14, 2021. 8

[5] Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatiotemporal state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. 8

[6] Bo Cheng, Yuhang Ma, Liebucha Wu, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Hico: Hierarchical controllable diffusion model for layout-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[7] A T Crooks, N W Gibin, and J J Heppenstall. A rule-based model of urban land use and economic growth. *International Journal of Geographical Information Science*, 24(2):311–331, 2010. 2

[8] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021. 3

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 6

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 6

[12] Hexiang Hu, Kelvin C.K. Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William W. Cohen, Ming-Wei Chang, and Xuhui Jia. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

[13] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2018. 6

[14] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. *arXiv preprint arXiv:2312.03606*, 2023. 2

[15] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023. 3, 5

[16] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[17] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. 8

[18] Liqin Liu, Bowen Chen, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Diverse hyperspectral remote sensing image synthesis with diffusion models. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16, 2023. 2

[19] Björn Lütjens, Brandon Leshchinskiy, Océane Boulais, Farrukh Chishtie, Natalia Díaz-Rodríguez, Margaux Masson-Forsythe, Ana Mata-Payerro, Christian Requena-Mesa, Aruna Sankaranarayanan, Aaron Piña, Yarin Gal, Chedy Raïssi, Alexander Lavin, and Dava Newman. Generating physically-consistent satellite imagery for climate visualizations. *IEEE*

Transactions on Geoscience and Remote Sensing, 62, 2024. 2

[20] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7465–7475, 2024. 2

[21] Li Pang, Xiangyong Cao, Datao Tang, Shuang Xu, Xueru Bai, Feng Zhou, and Deyu Meng. Hsigene: a foundation model for hyperspectral image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025. 2

[22] Junkil Park, Youhan Lee, and Jihan Kim. Multimodal conditional diffusion model using signed distance functions for metal-organic frameworks generation. *Nature Communications*, 16(1):34, 2025. 2

[23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2023. 5, 6

[24] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 5

[25] Moshe Pollak and David Siegmund. A diffusion process and its applications to detecting a change in the drift of brownian motion. *Biometrika*, 72(2):267–280, 1985. 3

[26] Y Ren, Y Lü, A Comber, B Fu, P Harris, and L Wu. Spatially explicit simulation of land use/land cover changes: Current coverage and future prospects. *Land Use Policy*, 80:324–336, 2019. 2

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[29] MS Roodposhti, J Aryal, and BA Bryan. A novel algorithm for calculating transition potential in cellular automata models of land-use/cover change. *Science of the Total Environment*, 674:290–303, 2019. 2

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2, 6

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[32] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021. 2

[33] Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021. 6

[34] Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36:62183–62223, 2023. 3

[35] Stability AI. Stable diffusion 3.5 technical report. https://stability.ai/news/stable-diffusion-3-5, 2024. 6

[36] Datao Tang, Xiangyong Cao, Xingsong Hou, Zhongyuan Jiang, Junmin Liu, and Deyu Meng. Crs-diff: Controllable remote sensing image generation with diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2

[37] Datao Tang, Xiangyong Cao, Xuan Wu, Jialin Li, Jing Yao, Xueru Bai, Dongsheng Jiang, Yin Li, and Deyu Meng. Aerogen: Enhancing remote sensing object detection with diffusion-driven data generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3614–3624, 2025. 2

[38] Kai Tang and Jin Chen. Changeanywhere: Sample generation for remote sensing change detection via semantic latent diffusion model. *arXiv preprint arXiv:2404.08892*, 2024. 2

[39] X Tong, H Liu, and Y Li. A statistical model for land use and land cover change prediction: A case study of urban growth in the beijing metropolitan area. *Environmental Modeling & Software*, 153:105418, 2022. 2

[40] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 8

[41] Haoyu Wang, Lei Zhang, Wei Wei, Chen Ding, and Yanning Zhang. Prompt-free conditional diffusion for multi-object image augmentation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pages 1945–1953, 2025. 2

[42] Ming Chen Wang and George Eugene Uhlenbeck. On the theory of the brownian motion ii. *Reviews of modern physics*, 17(2-3):323, 1945. 3

[43] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018. 2

[44] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5805–5813, 2024. 6

[45] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021. 6

[46] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Semantic change detection with asymmetric siamese networks. *arXiv preprint arXiv:2010.05687*, 2020. 6

[47] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, et al. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv:2403.11627*, 2024. 3

[48] Huan Ye, Jian Zhang, Shu Liu, Xinzhu Han, and Wenjun Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint*, arXiv:2308.06721, 2023. 6

[49] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18456–18466, 2023. 3

[50] Zhiping Yu, Chenyang Liu, Liqin Liu, Zhenwei Shi, and Zhengxia Zou. Metaearth: A generative foundation model for global-scale remote sensing image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[51] Qi Zang, Jiayi Yang, Shuang Wang, Dong Zhao, Wenjun Yi, and Zhun Zhong. Changediff: A multi-temporal change detection data generator with flexible text prompts via diffusion model, 2024. 2, 6

[52] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021. 6

[53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 6

[54] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023. 2

[55] Zhuo Zheng, Shiqi Tian, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Scalable multi-temporal remote sensing change data generation via simulating stochastic change process. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21818–21827, 2023. 2, 7

[56] Zhuo Zheng, Stefano Ermon, Dongjun Kim, Liangpei Zhang, and Yanfei Zhong. Changen2: Multi-temporal remote sensing generative change foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 6, 7

[57] Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. *arXiv preprint arXiv:2309.16948*, 2023. 3