ChangeBridge: Spatiotemporal Image Generation with Multimodal Controls for Remote Sensing —Supplementary Material—

1. Overview

The document supplements the main text with:

- Derivation of noise-asynchronous ChangeBridge in Sec. 2;
- Description of the LEVIR-CC dataset with coordinate annotations in Sec. 3;
- Additional ablation studies in Sec. 4;
- Performance on downstream change detection tasks in Sec. 5;
- Further qualitative results and visualizations in Sec. 6;
- Failure cases and limitations in Sec. 7.

2. Noise-Asynchronous ChangeBridge

Due to the strict theoretical definitions, the noise and drift in the diffusion bridge are intrinsically related. In this section, we present a theoretically rigorous version of the drift-asynchronous diffusion bridge, which uses asynchronous noise for both the foreground and background, along with asynchronous drift.

Asynchronous drift and noise. Given the drift map latent $\boldsymbol{z}_d$ that controls pixel-level drift during the diffusion process, we extend $\boldsymbol{z}_d$ as a diffusion map latent controlling both pixel-level drift and noise. This is achieved through the redefined drift coefficient, where $\tilde{m}_t(i,j) = m_t \cdot \boldsymbol{z}_d(i,j)$, with $m_t = \frac{t}{T}$. The corresponding pixel-wise noise coefficient is then defined as:

$$\begin{aligned}\tilde{\delta}_t(i,j) &= 2(\tilde{m}_t - \tilde{m}_t^2) \\ &= 2\left(m_t \cdot \boldsymbol{z}_d(i,j) - (m_t \cdot \boldsymbol{z}_d(i,j))^2\right).\end{aligned}$$

Then, we obtain the definition of asynchronous drift term $\tilde{m}_t(i,j)$ and asynchronous noise term $\tilde{\delta}_t(i,j)$:

$$\begin{cases}\tilde{m}_t(i,j) = \dfrac{t}{T} \cdot \boldsymbol{z}_d(i,j), \\ \tilde{\delta}_t(i,j) = 2\left(\dfrac{t}{T} \cdot \boldsymbol{z}_d(i,j) - \left(\dfrac{t}{T} \cdot \boldsymbol{z}_d(i,j)\right)^2\right).\end{cases} \tag{1}$$

Forward process. Now, the forward transition of the strictly defined asynchronous diffusion bridge can be expressed as:

$$q(\boldsymbol{z}_t \mid \boldsymbol{z}_b, \boldsymbol{z}_a) = \mathcal{N}\big(\boldsymbol{z}_t;\ (1 - \tilde{m}_t(i,j))\boldsymbol{z}_b + \tilde{m}_t(i,j)\boldsymbol{z}_a,\ \tilde{\delta}_t \boldsymbol{I}\big). \tag{2}$$

Reverse process. The reverse process is the same as the drift-asynchronous version; we incorporate the pixel-level diffusion magnitude latent $\boldsymbol{z}_d$ to adaptively reconstruct the heterogeneous dynamics of the foreground and background. The current reverse transition is as follows:

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

Training objective. The objective modifies both the drift term and the noise term as follows:

$$\mathcal{L}_{\text{asy}} = \mathbb{E}_{\boldsymbol{z}_a, \boldsymbol{z}_b, \boldsymbol{z}_c, \boldsymbol{\epsilon}}\Big[\big\|\tilde{m}_t(\boldsymbol{z}_a - \boldsymbol{z}_b) + \sqrt{\tilde{\delta}_t}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{z}_a, \boldsymbol{z}_c, \boldsymbol{z}_d)\big\|^2\Big] \tag{4}$$

Here, spatially adaptive drift and noise coefficient are introduced.

In practice, due to a) our task not requiring different random magnitudes for the foreground and background, and b) the asynchronous noise schedule $\tilde{\delta}_t$ making training convergence difficult, we adopt the drift-asynchronous version described in the main text.

3. LEVIR-CC Dataset with Coordinates

The original LEVIR-CC dataset consists of 10,077 pairs of bitemporal remote sensing images, each sized 256×256 pixels, accompanied by 50,385 sentences describing the differences between image pairs. The images capture various urban changes over time, providing a rich resource for change detection and captioning tasks.

To fit our setting of coordinate text as the condition, we add object counting and precise object-level coordinates to the sentences in the JSON file, leveraging the information from LEVIR-MCI [5]. The modifications are illustrated in Figure 1. These coordinates are highly accurate, and can be used to convert them into bounding boxes for composed bridge initialization.

4. Additional Ablation Studies

Different drift magnitudes. In this section, we perform an ablation study on the background drift magnitude γ_{bg} across three multimodal controls, with $\gamma_{fg} = 1.0$ set as the reference unit. As shown in Table 1, we observe significant performance gains when reducing γ_{bg} from 1.0 to the optimal value for each dataset. Specifically, for the semantic mask dataset (SECOND), the best performance is observed

Table 1. **Performance of different background drift magnitude** γ_{bg}. FID and IoU (%) are reported on WHU-CD, S2Looking, SECOND, and LEVIR-CC.

γ_{bg}	WHU-CD		S2Looking		SECOND		LEVIR-CC	
	FID ↓	IoU↑	FID ↓	IoU↑	FID ↓	mIoU↑	FID ↓	CosSim↑
1.0	56.24	71.87	80.38	73.21	70.45	59.04	46.73	0.69
0.9	48.52	73.04	76.64	75.30	68.25	64.77	42.84	0.73
0.8	**45.47**	**75.30**	**72.56**	**78.45**	65.31	69.98	**38.36**	**0.82**
0.7	47.84	74.48	74.29	77.68	**62.24**	**73.47**	41.79	0.77
0.6	49.69	73.56	75.81	76.92	66.17	70.55	40.56	0.70

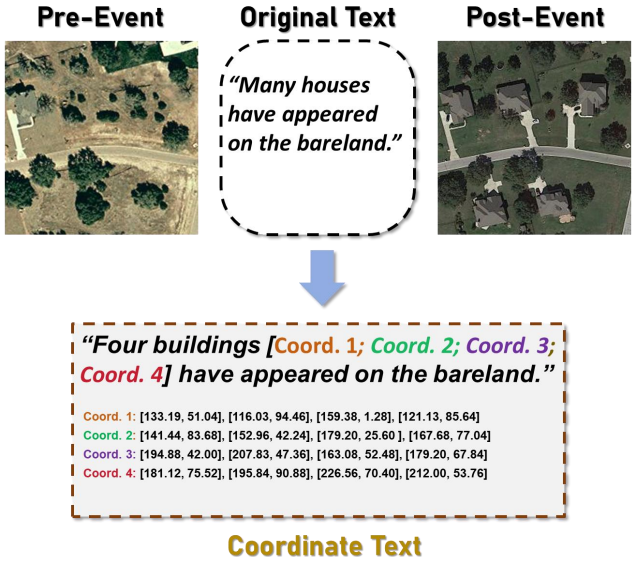

Figure 1. **LEVIR-CC dataset with coordinates**. we add object counting and precise object-level coordinates to the sentences.

Table 2. **Performance of our ChangeBridge with different backbones**. FID, IS, and IoU (%) are reported on WHU.

Backbone	FID ↓	IS ↑	IoU↑
SD1.5	76.81	4.85	65.29
UNet-based ChangeBridge	**45.47**	**5.88**	**75.30**
DiT	50.73	6.04	72.58
DiT-based ChangeBridge	**40.12**	**6.77**	**78.13**

at $\gamma_{bg} = 0.7$, resulting in a reduction in FID by 10.61 and an improvement in mIoU by 10.94%. For the instance layout datasets (WHU-CD and S2Looking), $\gamma_{bg} = 0.8$ achieves the best performance, reducing FID by 10.77 and 7.82, and improving IoU by 3.43% and 5.24%, respectively. For the coordinate text dataset (LEVIR-CC), the best result is at $\gamma_{bg} = 0.8$, with FID reduced by 8.37 and CosSim increased by 0.13, showing the most consistent performance improvement across all datasets.

Background consistency analysis. In this section, we evaluate whether the unchanged background regions are consistently preserved during cross-temporal generation. To this end, we report the Structural Similarity Index (SSIM) between generated and ground-truth post-event images specifically on background regions. As shown in Table 3, our method consistently achieves higher background SSIM across three representative multimodal settings, including instance layout (WHU-CD), semantic mask (SECOND), and coordinate text (LEVIR-CC). These results indicate better preservation of non-changing areas compared to existing methods.

Table 3. Background SSIM comparison under three conditions.

Dataset	Method	SSIM
Instance Layout	UNITE	0.824
	ControlNet*	0.510
	Ours	**0.879**
Semantic Mask	UNITE	0.786
	ControlNet*	0.802
	Ours	**0.825**
Coordinate Text	DreamBooth	0.862
	Instruct*	0.879
	Ours	**0.907**

Variants of ChangeBridge. We present our ChangeBridge implemented with different backbones, including the UNet and DiT variants, as shown in Table 2. For the SD1.5 baseline and DiT baseline, we incorporate pre-event and event controls into the denoising process, ensuring that the input and output remain consistent with our method. As shown in the table, compared to the baseline methods, we achieve significant improvements in performance, with our UNet variant reducing FID from 76.81 to 45.47, IS increasing from 4.85 to 5.88, and IoU rising from 65.29 to 75.30. Similarly, our DiT variant achieves a reduction in FID from 76.81 to 40.12, an increase in IS from 4.85 to 6.77, and an improvement in IoU from 65.29% to 78.13%. These results indi-

Table 4. **Evaluation of different fusion mechanisms of drift-aware denoising**. FID, IS, and IoU (%) scores are reported.

Conditional Fusion		FID ↓	IS ↑	IoU↑
SD1.5	Concat	50.95	5.63	73.26
	Cross-Attn	**45.47**	**5.88**	**75.30**
DiT	Concat	50.73	6.04	72.58
	FiLM	**40.12**	**6.77**	**78.13**

rectly validate the effectiveness of directly building cross-spatiotemporal diffusion for spatiotemporal image generation.

Fusion mechanisms of drift-aware denoising. We perform ablation for the fusion mechanism of the drift magnitude map z_d into the denoising network. As shown in Table 4, we evaluate the effectiveness of different fusion mechanisms, including concatenation (Concat), cross-attention (Cross-Attn), and FiLM modulation.

For the SD1.5 baseline, we observe that the cross-attention fusion mechanism achieves significant improvements, with FID decreasing by 5.48, IS increasing by 0.25, and IoU improving by 2.04%. For the DiT backbone, FiLM modulation outperforms the other methods, reducing FID by 10.61, increasing IS by 0.73, and improving IoU by 5.55%. These results help us select the most effective fusion mechanisms for drift-aware denoising.

5. More Results on Downstream Change Detection

In this section, we further evaluate the effectiveness of ChangeBridge as a data engine for downstream change detection tasks. Specifically, we focus on the most common binary change detection task. For baselines, we use the OpenCD toolkit [4] with four baselines: BiT [3], STANet [2], ChangeFormer [1], and ChangeStar [6]. Evaluation metrics include F1 score and IoU.

Table 5 shows the improvements of our methods across all baselines. Our ChangeBridge consistently enhances change detection performance across different models and datasets. On the WHU-CD dataset, as an augmentation technology, ChangeBridge improves the F1-score of STANet, BiT, ChangeStar, and ChangeFormer by +1.52%, +2.29%, +2.88%, and +2.38%, respectively, while the IoU gains range from +2.12% to +3.11%. Similarly, on the S2Looking dataset, ChangeBridge leads to F1-score increases of up to +2.27% and IoU improvements of up to +1.70%. Notably, ChangeFormer achieves the highest performance boost among all baselines, highlighting the adaptability of our augmentation to transformer-based architectures. These results confirm that ChangeBridge provides a strong data augmentation strategy, significantly benefiting change detection models by improving their ability to capture and recognize meaningful scene transformations.

Table 5. **Improvement of the data augmentation of our ChangeBridge for downstream change detection models**. F1 score (%) and IoU (%) are reported.

Method	WHU-CD		S2Looking	
	F1 ↑	IoU ↑	F1 ↑	IoU ↑
STANet	69.37	53.11	57.22	40.81
+Ours	**70.89**$_{+1.52}$	**55.42**$_{+2.31}$	**58.14**$_{+0.92}$	**42.30**$_{+1.49}$
BiT	83.98	72.39	62.74	46.94
+Ours	**86.27**$_{+2.29}$	**74.65**$_{+2.26}$	**64.05**$_{+1.31}$	**48.11**$_{+1.17}$
ChangeStar	84.16	74.02	66.30	49.75
+Ours	**87.04**$_{+2.88}$	**77.13**$_{+3.11}$	**68.57**$_{+2.27}$	**51.36**$_{+1.61}$
ChangeFormer	85.05	74.92	65.76	48.99
+Ours	**87.43**$_{+2.38}$	**77.04**$_{+2.12}$	**68.01**$_{+2.25}$	**50.69**$_{+1.70}$

5.0.1. Comparison with Change Synthesis Methods.

We evaluate ChangeBridge against existing change synthesis methods under two training settings: adaptation learning and zero-shot learning. In both cases, the models are evaluated on real test data, but the use of synthetic and real data in training differs. In adaptation learning, models are pre-trained on synthetic data and fine-tuned on real-world data, evaluating how the synthetic-to-real transfer enhances change detection. Zero-shot learning, on the other hand, uses only synthetic data for training, with models tested directly on real data without fine-tuning.

Adaptation performance. Table 6 shows the adaptation performance on the S2Looking dataset. Models are pre-trained on synthetic data and fine-tuned on real-world images. Results show that ChangeBridge outperforms all methods, especially those not using pre-event images.

Methods incorporating pre-event images outperform those without, as background variation due to factors like lighting, seasonal shifts, and human activity presents a major challenge in change detection. Using pre-event images allows models to handle these variations better, leading to more accurate change detection.

Among these methods, ChangeBridge achieves the best adaptation performance, improving the F1-score by +2.74 over the baseline and outperforming UNITE and ControlNet-IPA. This highlights the effectiveness of ChangeBridge's spatiotemporal diffusion approach in aligning spatial structures and ensuring semantically consistent change synthesis.

Zero-shot performance. We also evaluate zero-shot performance on the WHU-CD dataset, as shown in Table 7. In zero-shot learning, models are trained only on synthetic data and tested on real data. Despite this, ChangeBridge performs the best, surpassing all other methods.

When pre-event images are used, all methods improve, with ChangeBridge achieving an F1-score of 79.25, a +0.33 increase over ControlNet-IPA and +1.11 over UNITE. This improvement demonstrates the effectiveness of ChangeBridge's spatiotemporal diffusion approach in generating accurate changes, enhancing generalization in real-world scenarios.

These results emphasize the importance of pre-event-aware synthesis in both adaptation and zero-shot learning. ChangeBridge not only outperforms existing pre-event-based methods but also sets a new benchmark for zero-shot change detection performance. Its ability to align synthesized changes with spatiotemporal structures ensures reliable predictions even when models are trained solely on synthetic data.

Table 6. **Comparison of adaptation performance for change detection** with change synthesizing methods, with ChangeStar as the baseline on the S2Looking dataset. F1 score (%) is reported.

Method		Adaptation
Baseline		66.30
w/o Pre-Event Input	+Copy-Paste	66.70
	+Inpainting	66.30
	+Changen	67.10
	+Changen2	67.30
w/ Pre-Event Input	+UNITE	67.89
	+ControlNet-IPA	68.02
	+Ours	**69.04**

Table 7. **Comparison of zero-shot performance for change detection** with change synthesizing methods, with ChangeStar as the baseline on the WHU-CD dataset. F1 score (%) is reported.

Method		Zero-Shot
w/o Pre-Event Input	+Copy-Paste	3.80
	+Inpainting	21.80
	+Changen	26.60
	+Changen2	76.30
w/ Pre-Event Input	+UNITE	78.14
	+ControlNet-IPA	78.92
	+Ours	**79.25**

6. Further Qualitative Results

Example diversity. In this section, we present the diversity of generated examples, with sampling performed under the same conditions and pre-event images. As shown in Fig. 2, the generated samples vary significantly across examples while maintaining consistency with the corresponding controls and pre-event scenarios.

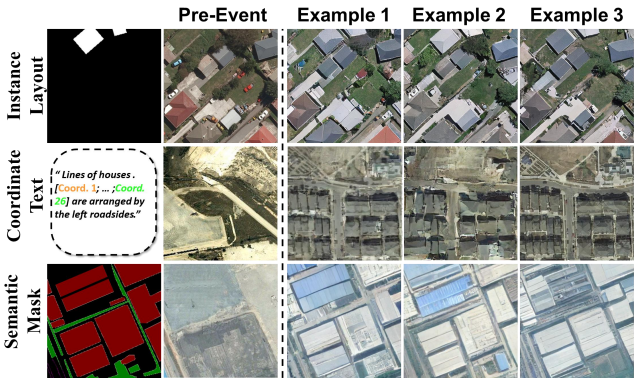

Figure 2. **Example Diversity of our ChangeBridge**, where sampling is performed under the same conditions and pre-event images.

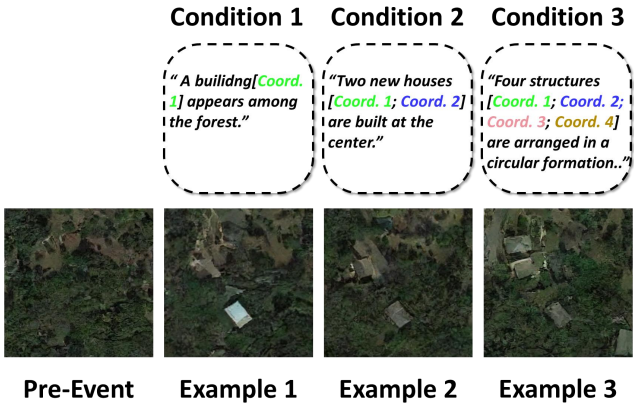

Figure 3. **Varying conditions on the same given scenes**, where sampling is performed on the same pre-event image with different conditional contexts. This shows the interactive capability of ChangeBridge for land planning.

Varying conditions on the same scenes. To demonstrate the interactive capability of ChangeBridge for land planning scenarios, we enable users to provide different control contexts for the same scene. As shown in Figure 3, we apply varying coordinate texts on the same pre-event image, allowing users to adjust the conditional context. This results in diverse post-event scenarios while maintaining the spatial and temporal consistency, highlighting the flexibility and interactivity of ChangeBridge in supporting dynamic land planning tasks.

Visualization of pure cross-temporal generation. In this section, we present pure cross-temporal generation samples without additional event controls on the WHU-CD dataset. This is achieved by sampling from a zeroed control latent. As shown in Figure 4, this demonstrates the model's capability in spatiotemporal generation modeling.

More comparison and synthesized examples. We present additional comparisons with multi-condition methods and

Figure 4. **Visualization of pure cross-temporal generation** without additional controls. This demonstrates the model's capability in spatiotemporal modeling.

change generation methods, including instance-layout generation in Figure 6, semantic-mask generation in Figure 7, and coordinate-text generation in Figure 8.

7. Failure Cases and Limitations

Failure cases. In this section, we analyze challenging scenarios where the proposed method may fail. As shown in Fig. 5, road regions remain a typical failure case. Due to their thin and elongated structures, these regions are highly sensitive to minor appearance variations, which may lead to broken connectivity or distorted artifacts in the generated results.

Figure 5. Examples of failure cases, where thin structures such as roads are difficult to preserve.

These observations indicate that preserving structural continuity and reducing domain discrepancy remain key challenges. Future work may incorporate topology-aware constraints and artifact-aware strategies to mitigate the synthetic–real domain gap.

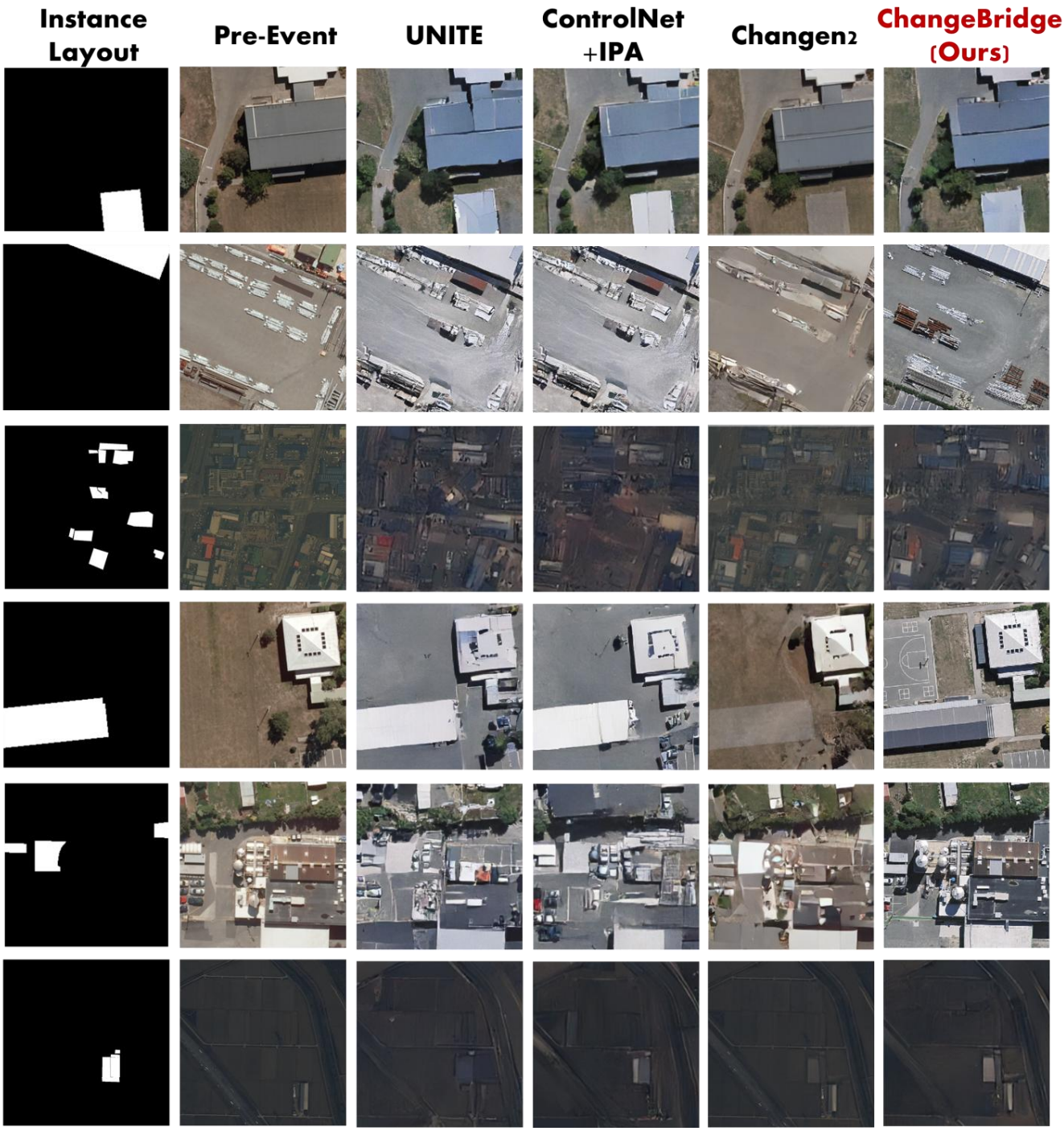

Figure 6. Comparison of instance-layout generation results.

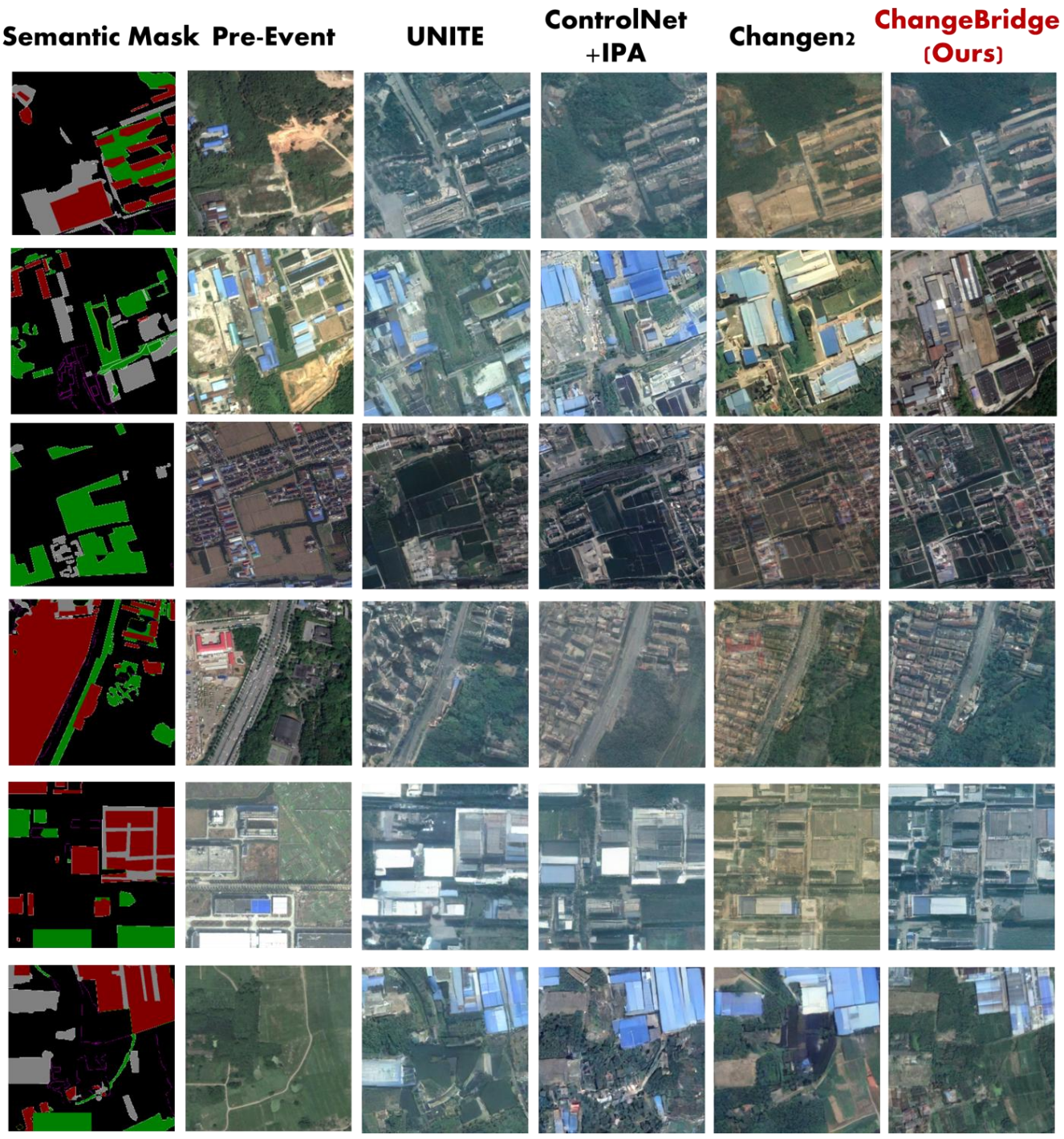

Figure 7. Comparison of semantic-mask generation results.

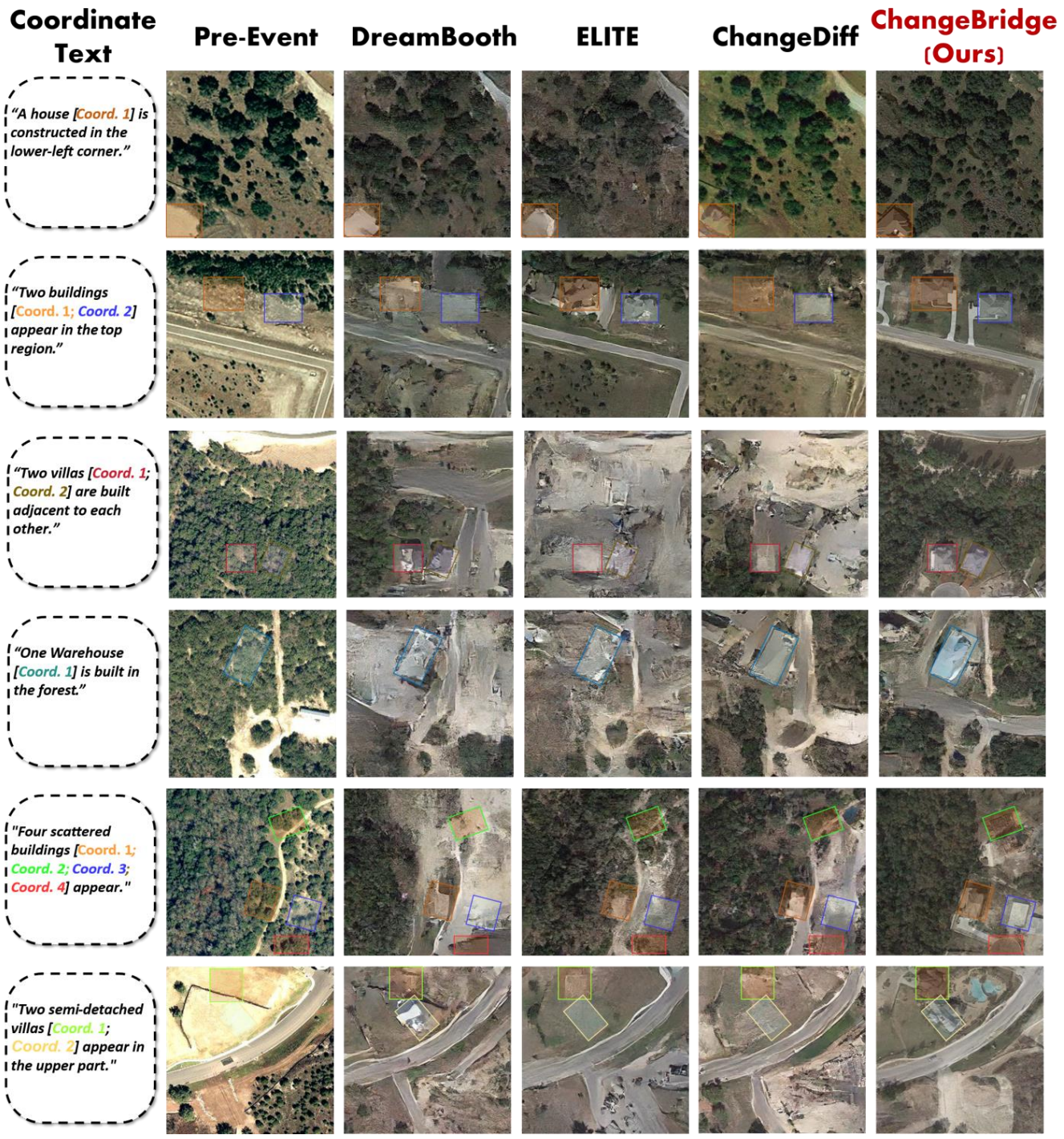

Figure 8. Comparison of coordinate-text generation results.

References

[1] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022. 3

[2] 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. 3

[3] 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. 3

[4] Kaiyu Li, Jiawei Jiang, Andrea Codegoni, Chengxi Han, Yupeng Deng, Keyan Chen, Zhuo Zheng, Hao Chen, Zhengxia Zou, Zhenwei Shi, et al. Open-cd: A comprehensive toolbox for change detection. *arXiv preprint arXiv:2407.15317*, 2024. 3

[5] Chenyang Liu, Keyan Chen, Haotian Zhang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. Change-agent: Towards interactive comprehensive remote sensing change interpretation and analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1

[6] Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15193–15202. IEEE, 2021. 3