Diffusion-based Semantic Image Synthesis from Sparse Layouts

Yuantian Huang, Satoshi Iizuka, Kazuhiro Fukui

University of Tsukuba

1 Introduction

Background

- **Previous approaches** on the task of Semantic Image Synthesis
  - Use **detailed and precise semantic layouts**, while the quality of the results is highly dependent on the accuracy of the input layouts.
  - However, it is quite **challenging** for real users to create highly detailed and accurate semantic layouts in practice.

Figure 1: Examples from previous research*.

*: SPADE: Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization, 2019.

➢ Our Goal

- **Proposed approach** → Synthesize images from **sparse and intuitive semantic layouts.**

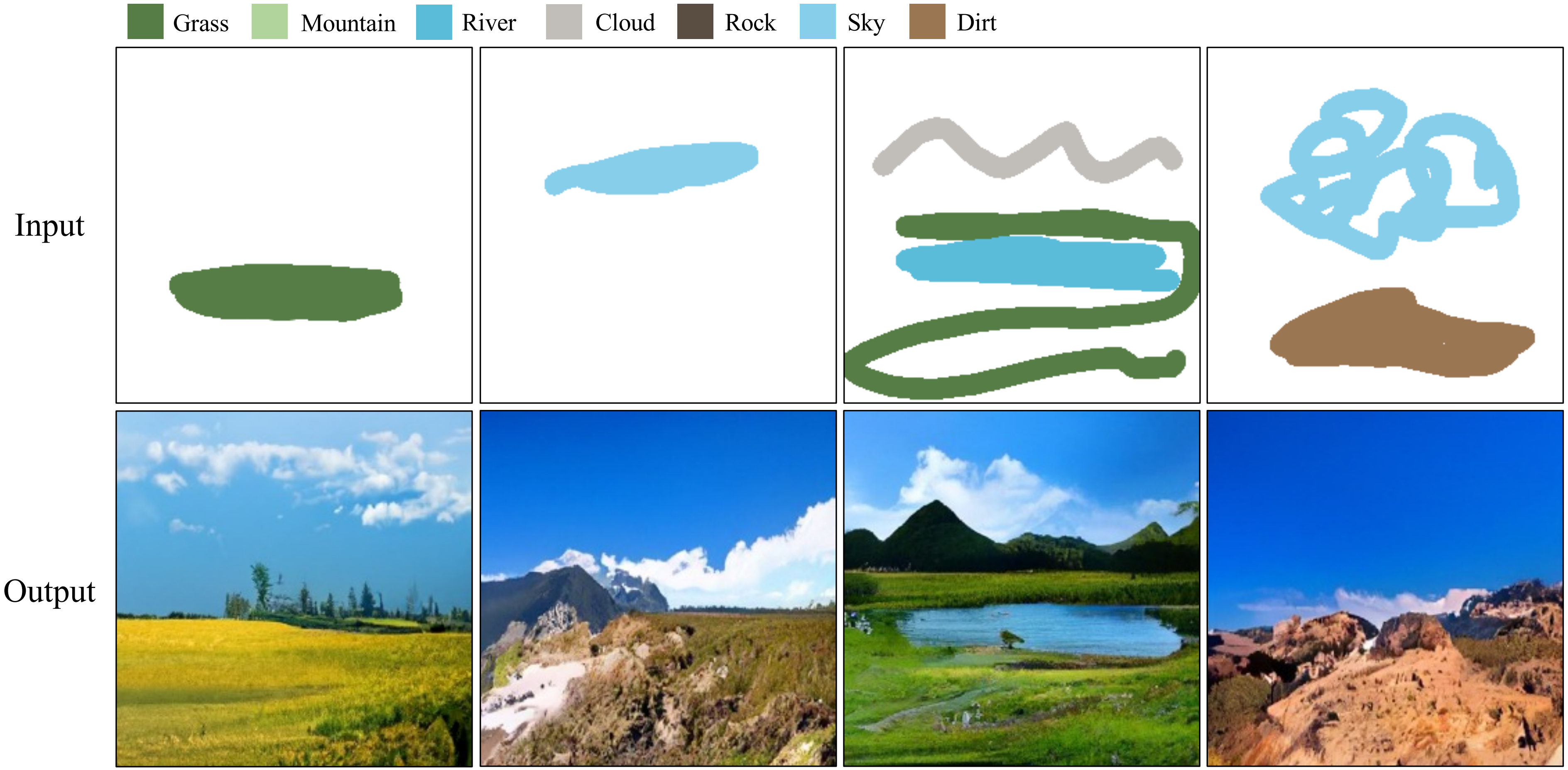

Figure 2: Teaser for our proposed method.

CVLAB

筑波大学
University of Tsukuba

2 Proposed Framework

Overview

1. **A random masking process** that tries to simulate actual user input, improving generation quality in practical applications during the inference stage.

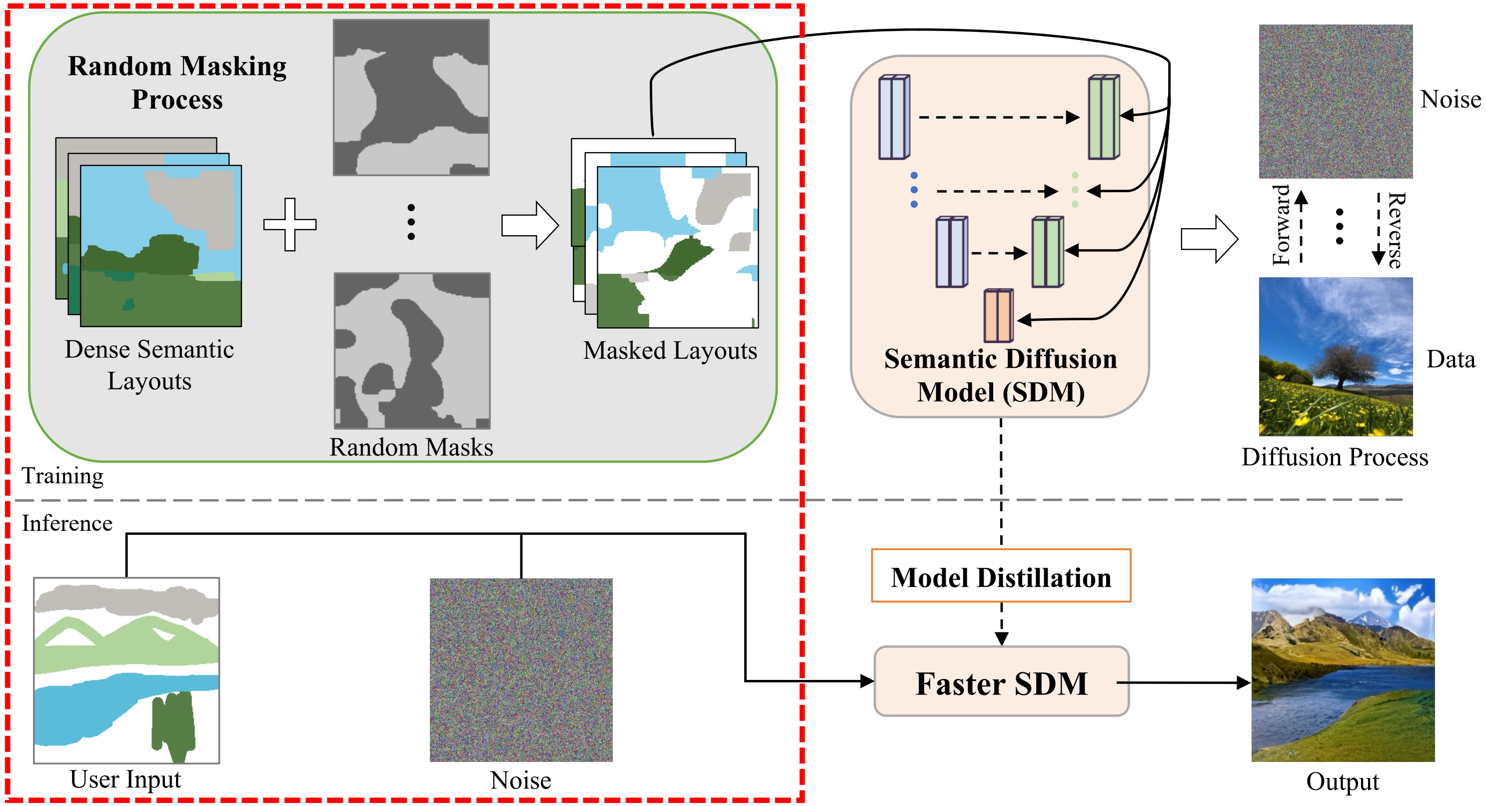

Figure 3: Structure of Proposed Framework.

➢ Overview

2. **A diffusion-based generator*** that we found to be most suitable for our masking process while also surpassing previous GAN-based models in generation quality.

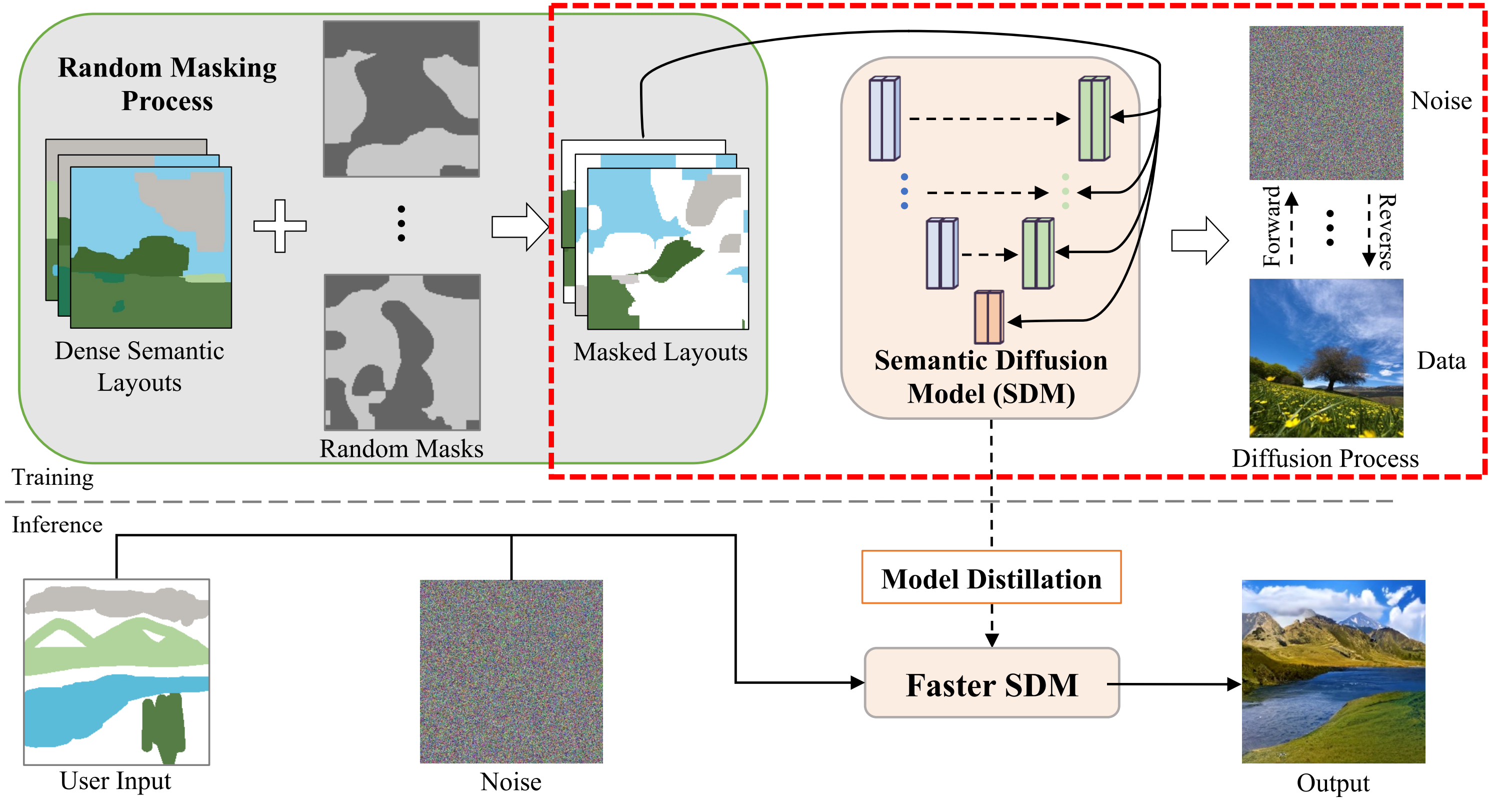

Figure 3: Structure of Proposed Framework.

*: SDM: Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L. and Li, H.: Semantic Image Synthesis via Diffusion Models (2022).

➢ Overview

3. **A progressive model distillation* process** that significantly reduces diffusion steps during the inference stage, making our framework interactive and broadly applicable.

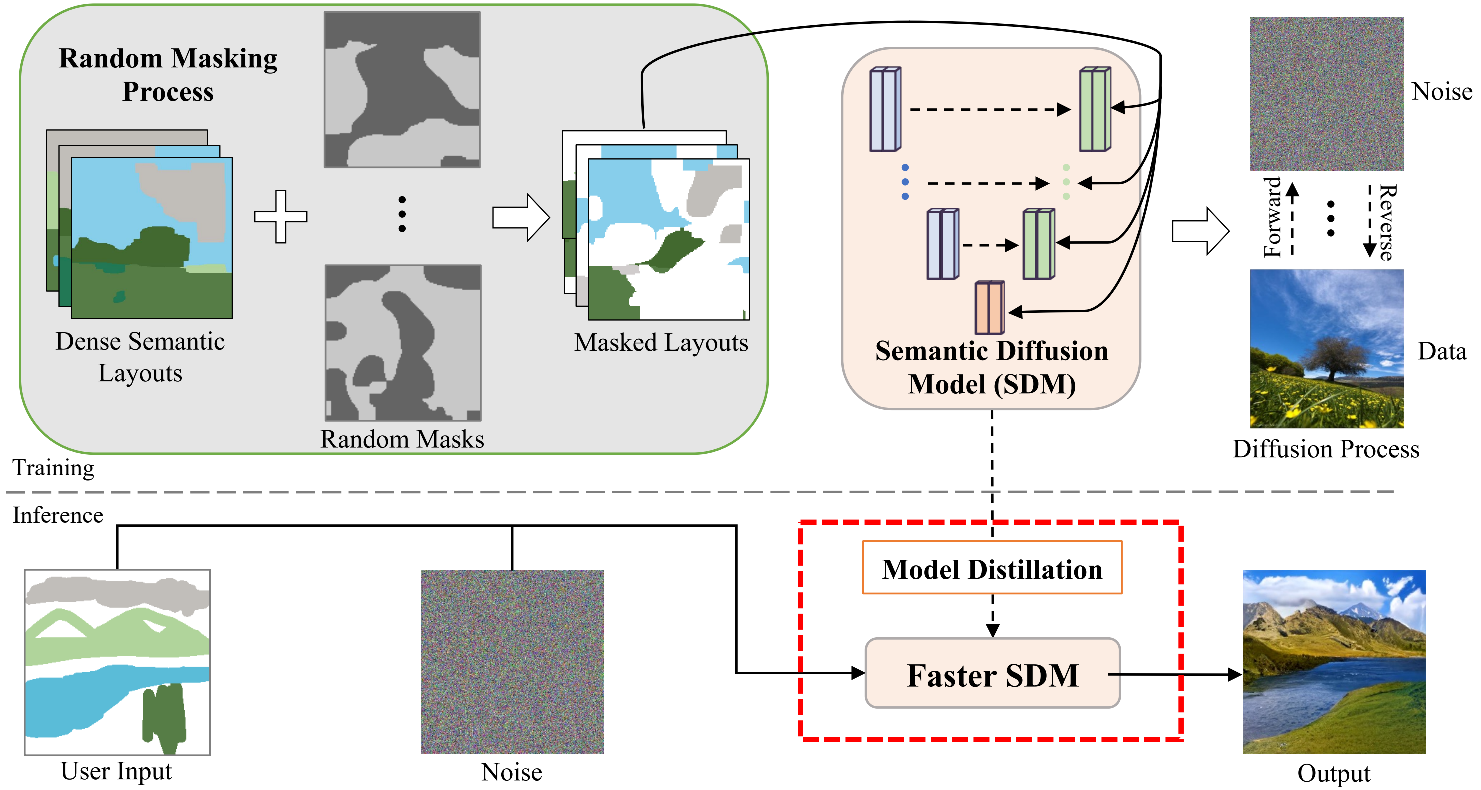

Figure 3: Structure of Proposed Framework.

*: Salimans, T. and Ho, J.: Progressive Distillation for Fast Sampling of Diffusion Models (2022).

➢ Random Masking Strategy

- We propose to simulate human-authored semantic layouts during the training using a well-designed random masking strategy called **Class-wise Random Patterns that**
  - generates **random patterns to mask a certain percent** of the semantic layouts.
  - generates **different masks for each label class** to avoid biased learning.

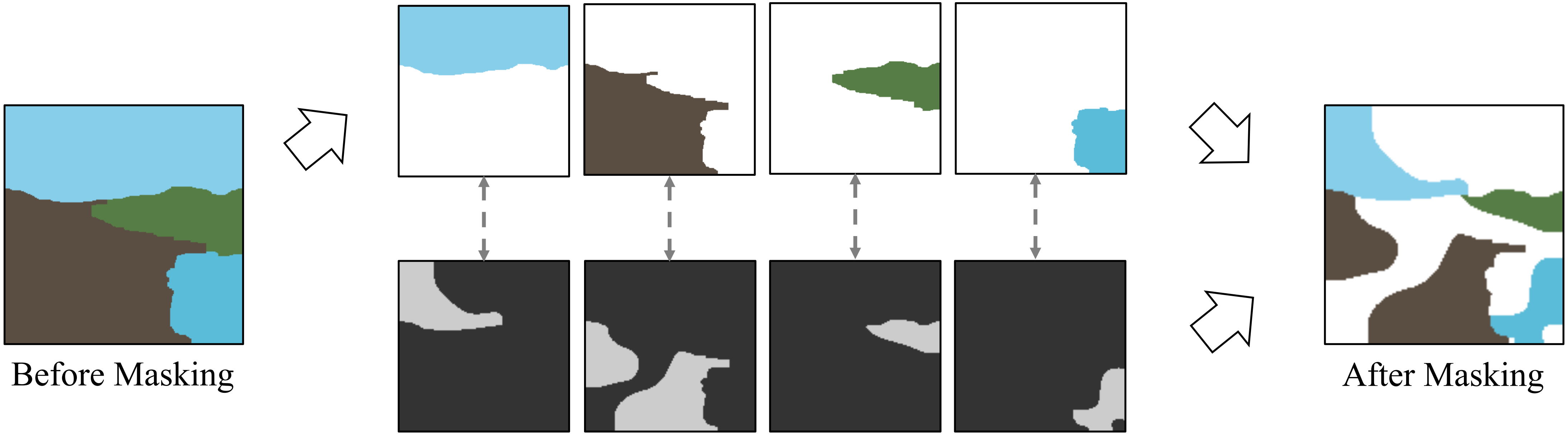

Figure 4: Example of our random masking strategy, different masks are generated for each label class.

3 Results

➢ Baseline Models

- We chose three existing models as our baseline models:
  - Semantic image synthesis with spatially-adaptive normalization. (SPADE).
  - You Only Need Adversarial Supervision for Semantic Image Synthesis. (OASIS)
  - Image Synthesis with Semantic Region-Adaptive Normalization. (SEAN)

- OASIS: Sushko, V., Schönfeld, E., Zhang, D., Gall, J., Schiele, B. and Khoreva, A.: You only need adversarial supervision for semantic image synthesis. (2020).
- SPADE: Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization, 2019.
- SEAN: Zhu, P., Abdal, R., Qin, Y. and Wonka, P.: SEAN: Image Synthesis with Semantic Region-Adaptive Normalization. (2020).

Quantitative Comparison

- Fréchet Inception Distance (FID)
  - Fréchet distance between two multidimensional Gaussian distributions, which captures the perceptual similarity of generated images to real ones.
  - The lower, the better.
  - As shown in Table 1, our approach outperforms baseline models in terms of generation quality.

	SPADE	SEAN	OASIS	ours
FID↓	57.82	148.32	44.47	**38.37**

Table 1: Quantitative comparison with baseline models.

➢ Qualitative Comparison

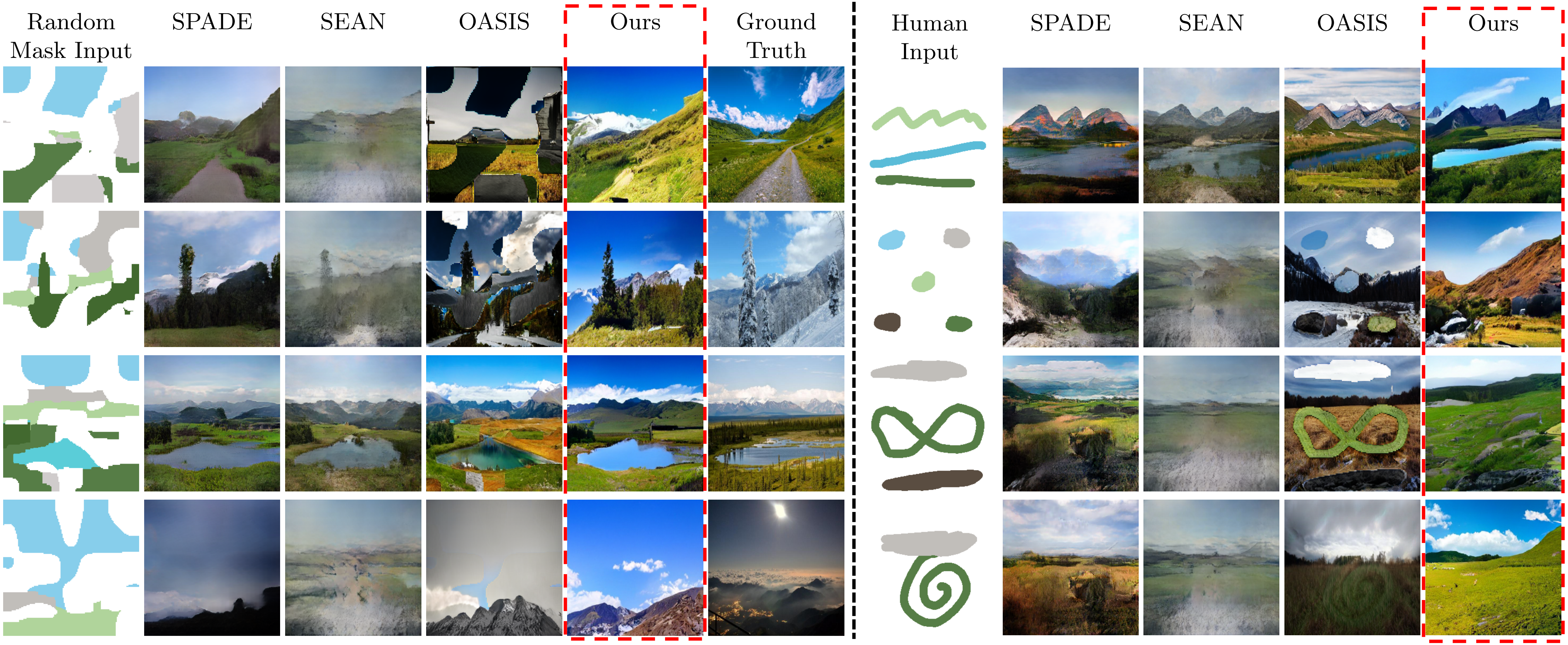

Figure 5: Qualitative comparison of generation results from both random mask input and actual human input.

➢ Example of Interactive Editing

➢ Our Contribution

- **Diffusion-based Semantic Image Synthesis from Sparse Layouts.**

  1. A well-designed masking strategy that simulates human-authored sparse layouts, avoiding the challenging task of producing detailed semantic layouts.

  2. A diffusion-based generator tailored to our masking design, which outperforms existing GAN-based models in terms of generation quality.

  3. An additional model distillation process makes our framework more interactive and applicable for practical use.

Thank You For Listening!