16726 Spring 25 Final Project Proposal

Keling Yao[†] Yinghao Zhang[†] Silong Yong[†]
[†]: Equal contribution
Robotics Institute, School of Computer Science
{kennyy, yinghaoz, silongy}@andrew.cmu.edu

1. Overview

In the final project of 16726, we plan to propose a refinement method on MultiDiffusion [3]. MultiDiffusion is a training-free paradigm for controllable image generation that unifies multiple diffusion processes through shared constraints, enabling seamless synthesis across overlapping regions or semantically distinct areas. By optimizing an objective that reconciles diverse denoising directions from a pre-trained diffusion model, it supports tasks such as panorama creation and region-based text-to-image generation. MultiDiffusion achieves this without modifying model weights, setting it apart from earlier methods relying on fine-tuning (e.g., SpaText [2], Make-A-Scene [4]) or task-specific pipelines like Blended Latent Diffusion [1].

While MultiDiffusion has shown promise in panorama generation and region-based text-to-image synthesis individually, combining these two capabilities in a unified setting introduces new challenges. Our goal is to generate **panoramic images with region-level controllability**, allowing users to specify fine-grained content within a wide, high-resolution canvas.

In this project, we aim to make the following contributions:

- A unified, training-free pipeline for generating panoramic images with region-level control.
- A refinement framework to improve semantic alignment and coherence across spatial regions.

2. Problem Statement

Input: A text prompt and region-level spatial masks, along with the desired resolution (e.g., 512 × 2048).

Output: A coherent high-resolution image with controllable layout fidelity and seamless transitions.

The integration of region-based control with panoramic generation introduces new challenges. In particular, semantic consistency across regions becomes more difficult to maintain. For instance, as shown in Figure 1, there are inconsistency in the middle of the "green field", despite being conditioned on a shared region mask. This is due to independently evolving diffusion paths that can diverge semantically despite shared pixel overlaps. In this project, we aim to address these challenges by exploring refinement techniques that enhance cross-region coherence.

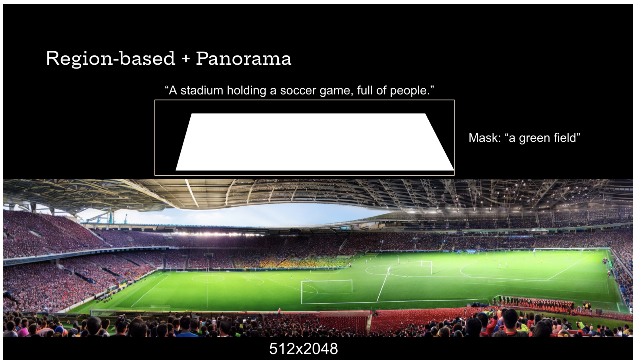
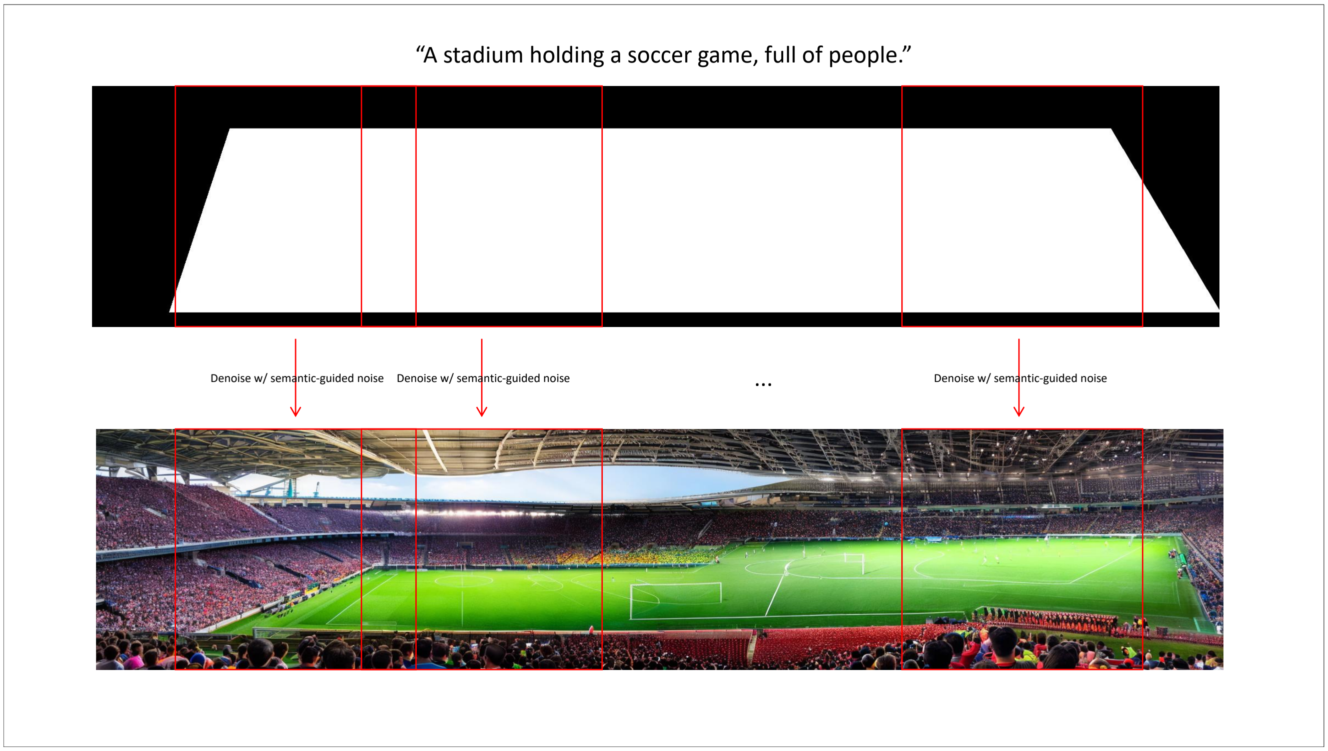


Figure 1. Results from MultiDiffusion. Semantic inconsistency when combining region-based constraints with panorama generation.

3. Methodology

As shown in Figure 1, existing method lacks the capability of context-awareness when generating new regions for the panorama. Such a design choice makes the generation in the late iteration overwrite the previous generated content without carefully blending the edges of the region. The overall idea can be found in Figure 2. For the semantic-guided noises, our initial plan for its design can be categorized as two approaches. First, we plan to investigate ways of incorporating existing generated region into the current iteration, providing useful contextual information to help the model generate consistent regions. Second, we plan to investigate the attention matrix in the textual encoder as well as the model backbone to see how the semantics is grounded to the generation process, which can be modified and utilized accordingly for our purpose. The semantic-guided noise in Figure 2 will consists of information derived from the above mentioned approaches. We expect such in-depth investiga-



"A stadium holding a soccer game, full of people."
Denoise w/ semantic-guided noise
Denoise w/ semantic-guided noise
...
Denoise w/ semantic-guided noise

Figure 2. Results from MultiDiffusion. Semantic inconsistency when combining region-based constraints with panorama generation.

tion for context- and semantic-aware generation could help with generating panorama with correct and consistent semantic regions.

4. Action Plan

Timeline:

- **Mar. 25 - Mar. 31**: Literature review and background study.
- **Apr. 1 - Apr. 4**: Run wide baseline experiments on MultiDiffusion, including region-based generation and panorama generation and combining both.
- **Apr. 5 - Apr. 19**: Design and implement the refinement framework. Run experiments to evaluate, analyze, and compare with baselines.
- **Apr. 20 - Apr. 22**: Prepare for the presentation.
- **Apr. 23 - Apr. 30**: Make some improvements according to the feedback from the presentation. Build the final report website.

Resources: The project will be conducted using the existing MultiDiffusion codebase. We will also leverage the pre-trained Diffusion model. The computational resources will include multiple NVIDIA A6000 GPUs.

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, 42(4):1–11, 2023. 1

[2] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 18370–18380. IEEE, 2023. 1

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation, 2023. 1

[4] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors, 2022. 1