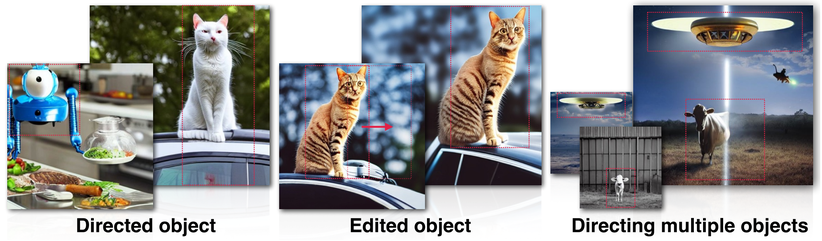
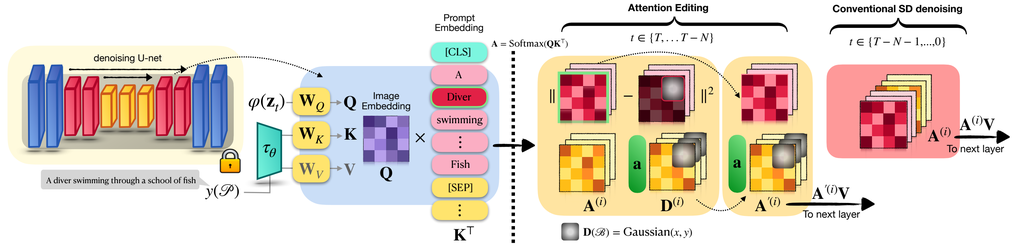
Directed Diffusion:
Direct Control of Object Placement through Attention Guidance

Wan-Duo Kurt Ma [1] Avisek Lahiri [2] J.P. Lewis [3*] Thomas Leung [2] W. Bastiaan Kleijn [1,2]

[1]Victoria University of Wellington [2]Google Research [3]NVIDIA Research [*]Work done at Google Research





Introduction

Directed Diffusion (DD) provides easy high-level positional control over multiple objects, while making use of an existing pre-trained denoising diffusion model and maintaining a coherent blend between the positioned objects and the background through cross-attention map editing without neural network training or finetuning.

Contribution

- **Storytelling**: DD provides control over the positioning of multiple objects.
- **Compositionality**: DD provides a direct approach to "compositionality" by providing explicit positional control.
- **Consistency**: The positioned objects seamlessly and consistently fit in the environment, rather than appearing as a splice from another image with inconsistent interaction.
- **Simplicity**: DD allows the user to control the desired locations of objects simply by specifying approximate bounding boxes. It requires only a few lines to implement our core method.

Background

The overall position and shape of a synthesized object (e.g., the cat) appears near the beginning of the denoising process, while the final denoising steps do not change this overall position but add details that make it identifiable as a particular object.

Figure 1. (Left) Final VAE reconstruction. (Right) Cross-attention map associated with the word in prompt "cat" from the beginning to the end of denoising process.

Results

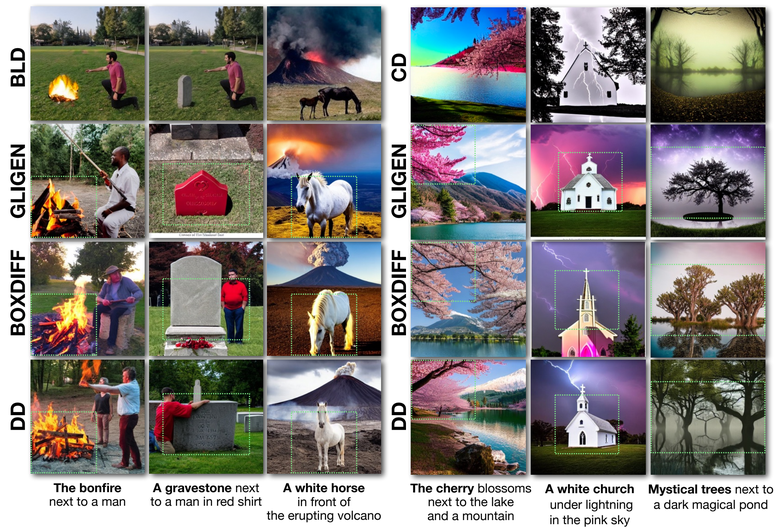


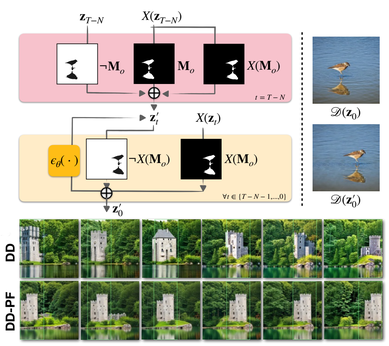
Figure 2. Various results compared to BOXDIFF (Xie et al., 2023), GLIGEN (Li et al., 2023), BLD (Avrahami, Fried, and Lischinski, 2022), and CD (Liu et al., 2022)

Methodology

Our lightweight optimization objective seeks to find the best weighed combination of "trailing" cross-attention maps at time t, by means of a weight vector $\mathbf{a}_t \in \mathbb{R}^{77-|\mathcal{P}|-1}$ for the "directed" prompt maps $\mathbf{A}^{(i)}_{t-1}$. The on-line optimization attempts to match the corresponding target maps $\mathbf{D}^{(i)}$ with the associated i_{th} prompt word:

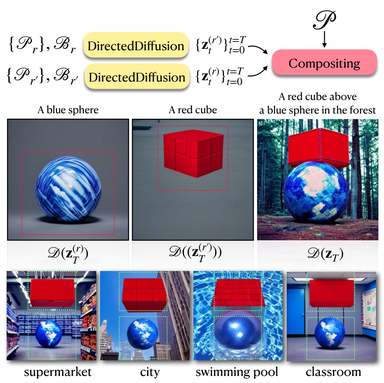
$$\mathcal{L}_{\mathbf{a}_t} = \sum_i \left\| \mathbf{A}^{(i)}_{t-1} \left(\mathbf{A}^{(|\mathcal{P}|+1:77)}_t \cdot \text{Diag}(\mathbf{a}_t) \right) - \mathbf{D}^{(i)} \right\|^2$$

Application: Placement Finetuning



In some cases the artist may wish to experiment with different object positions after obtaining a desirable image. However, when the object's bounding box is moved DD may generate a somewhat different object. Our placement finetuning (PF) method addresses this problem, allowing the artist to immediately experiment with different positions for an object while keeping its identity, and without requiring any model finetuning or other optimization.

Application: Scene Compositing



Prompts involving several objects often fail in T2I models, especially those based on CLIP. Our DD pipeline supports the direction of multiple objects by means of multiple bounding boxes. However, in practice the results are unreliable when the number of bounding boxes is more than two. We resolve this problem by additional editing operations.

Extension

Our work has been extended to video generation: **Trailblazer: trajectory control for diffusion-based video generation**. Please visit our site for more information: https://hohonu-vicml.github.io/Trailblazer.Page/

Directed Diffusion TrailBlazer