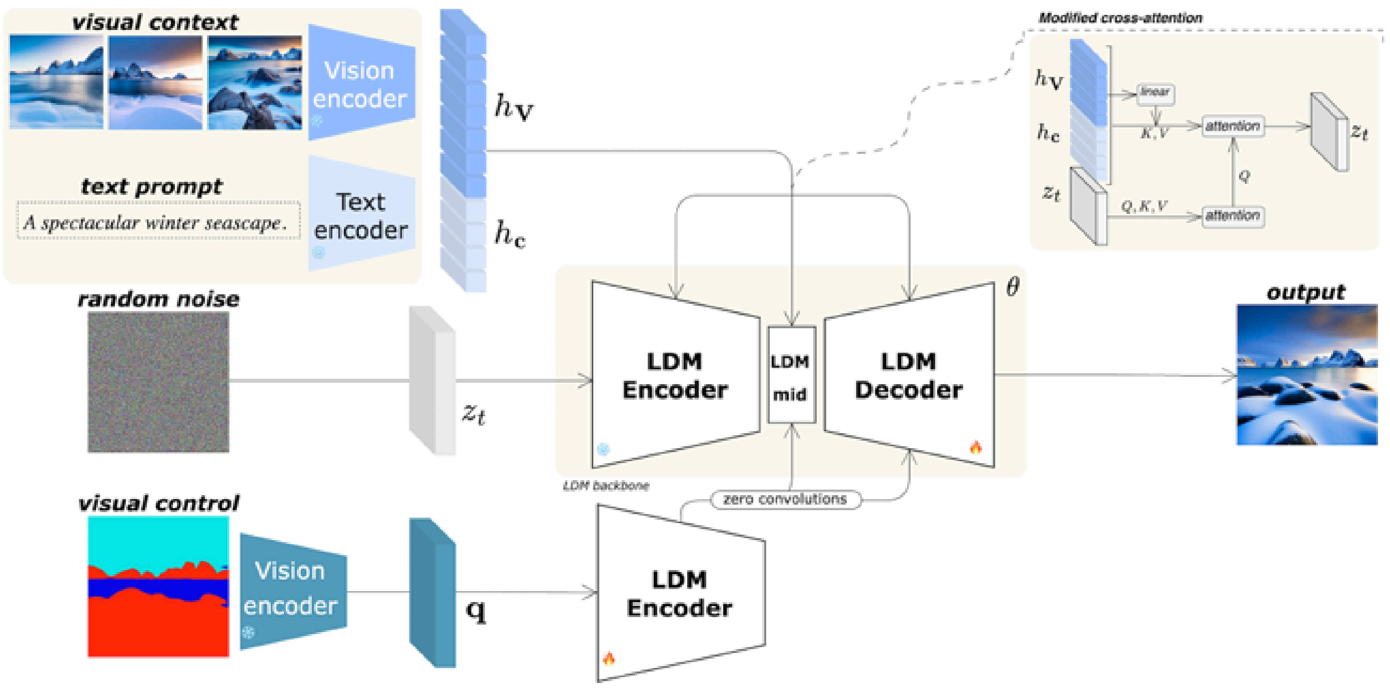
# Context Diffusion: In-Context Aware Image Generation

## Introduction

- **Objective**: Develop an in-context learning framework for image generation using visual context alone or with minimal text prompts.
- **Problem Statement**: Traditional models rely heavily on text prompts, limiting flexibility when only visual examples are available.
- **Solution**: Context Diffusion enables few-shot image generation using multiple visual examples or no prompts, enhancing adaptability across in-domain and out-of-domain tasks.




## Proposed Methodology

### 1. Diffusion Model Overview

The Context Diffusion model is built on a denoising diffusion framework, which generates images by progressively refining noisy data. The objective function is given by:

$$L = E_{z,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - f_\theta(z_t, t, c)\|^2\right] \qquad (1)$$

where:

- $z_t$ represents the noisy data representation at timestep $t$,
- $\epsilon$ is Gaussian noise,
- $c$ is a conditioning variable that includes text and/or visual embeddings,
- $f_\theta$ is the model function, parameterized by $\theta$.

### 2. Architecture Components

- **Text Prompt Encoding**: CLIP's text encoder generates semantic embeddings from text prompts.
- **Visual Context Encoding**: Visual context images are encoded and averaged to form a unified embedding.
- **Cross-Attention Layers**: Combines text and visual embeddings, allowing the model to attend to both or rely on visual context alone.
- **Layout Control with Query Image**: The query image guides the structure, maintaining spatial coherence through a layout-preserving approach.

### 3. Cross-Attention Mechanism

To integrate both text and visual context, Context Diffusion employs a cross-attention mechanism. The cross-attention operation is given by:

$$z_t = z_t + \text{CrossAtt}(Q = z_t, K = V = [h^c, h^V]) \qquad (2)$$

where:

- $h^c$ represents text embeddings,
- $h^V$ represents visual embeddings.

By concatenating text and visual embeddings, the model achieves high-quality outputs even without text prompts, as the model learns to manage various contexts effectively.

## Results

### In-Domain Tasks:

- Context Diffusion achieves high fidelity to visual context with color and style consistency.
- Demonstrates lower FID and RMSE scores compared to models relying heavily on text prompts.

### Out-of-Domain Tasks:

- Effective adaptability for sketch-to-image tasks and similar applications.
- Outperforms traditional models in human evaluation, achieving a preference rate above 55%.

## Conclusion

- **Contribution**: Context Diffusion introduces a flexible image generation approach that minimizes dependency on text prompts.
- **Significance**: Expands image generation capabilities for applications requiring high fidelity to visual context alone.

## References


- Ivona Najdenkoska, Animesh Sinha, Abhimanyu Dubey, Dhruv Mahajan, Vignesh Ramanathan, Filip Radenovic. "Context Diffusion: In-Context Aware Image Generation." Meta GenAI, University of Amsterdam.