Context Diffusion: In-Context Aware Image Generation

Project Overview

The **Context Diffusion Model** generates images by combining:

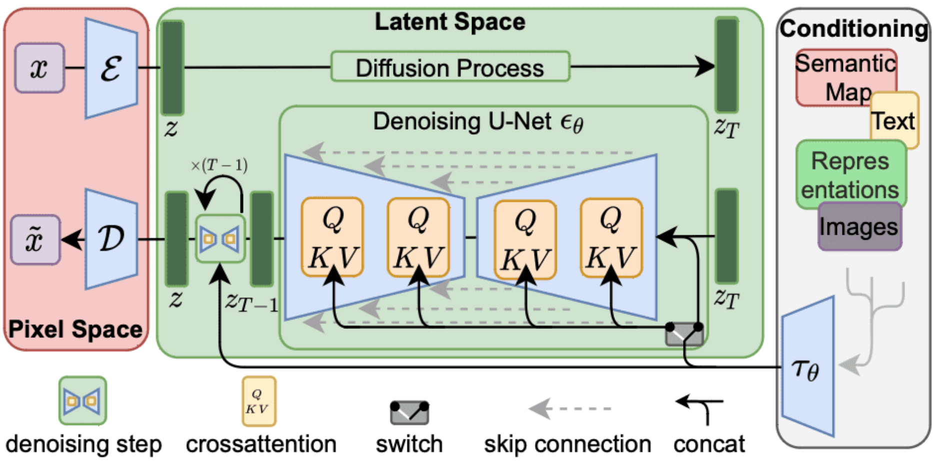
- **Query Image**: Defines the layout/structure.
- **Context Images**: Provide style and texture.
- **Text Prompt**: Adds semantic guidance.

Using CLIP embeddings and a Latent Diffusion Model (LDM), it outputs images that align with a given structure, style, and prompt description.

Key Features

- **CLIP-Based Encoding**: Embeds text and context images.
- **Latent Diffusion Model (LDM)**: Generates images using the combined embeddings.
- **Dynamic Image Normalization**: Calculates mean and standard deviation per image for flexible normalization.

Model Workflow



Working Mechanism

1. **Diffusion Model Objective**: The model's denoising objective seeks to generate samples by progressively removing noise. Given noisy data z_t at time t, the loss function is:

$$L(\theta) = E_{z,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - f_\theta(z_t,t,c)\|_2^2\right]$$

where c is the conditioning input.

2. **Conditioning Information**: The model combines multiple conditioning sources $y = (c, V)$, with V representing the set of context images:

$$V = [v_1, v_2, \ldots, v_k]$$

Each context image v_i is encoded into embeddings h_{v_i}, and the combined visual embedding h_V is computed by:

$$h_V = \sum_{i=1}^{k} h_{v_i}$$

3. **Text and Visual Context Encoding**: Using separate CLIP models, the text prompt and context images are embedded as follows:

$$h_c = \{h_c^0, h_c^1, \ldots, h_c^{N_c}\} = f_{\text{text}}(c)$$
$$h_{v_i} = \{h_{v_i}^0, h_{v_i}^1, \ldots, h_{v_i}^{N_v}\} = f_{\text{img}}(v_i)$$

4. **Cross-Attention Mechanism**: The combined embedding from the text and visual context, $[h_c, h_V]$, conditions the diffusion model via cross-attention, updating z_t:

$$z_t = z_t + \text{CrossAtt}(Q = z_t, K = V = [h_c, h_V])$$

5. **Final Objective with Query Conditioning**: Integrating both context and query image conditioning, the loss function becomes:

$$L(\theta) = E_{z,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - f_\theta(z_t,t,y,q)\|_2^2\right]$$

where q represents the query image's structural guidance.