Context Diffusion: In-Context Aware Image Generation

Introduction

The **Context Diffusion** framework addresses limitations in traditional image generation models that struggle to incorporate visual context without relying heavily on text prompts. This work introduces a diffusion-based model that utilizes both visual and textual conditioning, effectively enabling in-context learning for image generation. The model is designed to work in few-shot and zero-shot settings, allowing it to perform well even when only a few visual examples are provided, or no textual prompt is available. This approach opens up new possibilities for domains such as media content creation and personalized image generation, where textual prompts may be limited or unavailable.

Methodology

1. Diffusion Model Overview

Diffusion models generate images by progressively denoising a random noise vector, guided by both visual and textual context information. In Context Diffusion, the model minimizes the following objective to learn the distribution of images conditioned on given context:

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

where:

- z_t represents the noisy image state at timestep t,
- ϵ is the noise term sampled from a normal distribution,
- c is the context variable that includes text and/or visual embeddings,
- f_θ is the model's function, parameterized by weights θ.

The diffusion model learns to iteratively refine noisy data by leveraging context from images and/or text, recovering a high-quality image from a pure noise starting point.

2. Architecture of Context Diffusion

Context Diffusion extends the standard Latent Diffusion Model (LDM) by introducing dual-conditioning mechanisms: text-based and visual context-based conditioning. Key components include:

- **Prompt Encoding**: Text prompts are encoded via a CLIP text encoder, generating a fixed-size embedding that provides semantic context.
- **Visual Context Encoding**: Visual context, potentially including multiple images, is encoded using a pre-trained CLIP image encoder. Multiple visual context images are averaged to create a unified embedding that encapsulates color, style, and spatial layout.
- **Cross-Attention Layers**: Text and visual embeddings are combined and processed by cross-attention layers, enabling the model to attend to both contexts. This mechanism allows the model to selectively emphasize parts of the visual context based on the text prompt or purely rely on visual examples if the text prompt is absent.

- **Query Image and Layout Control**: A separate "query image" acts as a layout guide. This layout-preserving approach ensures the spatial coherence of generated images by conditioning the generation process on the structure of the query image, similar to ControlNet.

3. Training and Optimization

The training objective of Context Diffusion is adapted to jointly handle text and visual conditioning, integrating both contexts during each denoising step. The adapted loss function is defined as:

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

where:

- y represents the combined text and visual context,
- q represents the query image that helps preserve layout during generation.

Training is performed on a range of tasks, including in-domain and out-of-domain tasks, to ensure that the model generalizes well to new scenarios.

Implementation Details

The code is implemented using `PyTorch` along with `transformers` and `diffusers` libraries from Hugging Face. Key aspects of the code implementation include:

- **Device Selection**: The model is configured to use a GPU if available, enhancing computational efficiency.
- **Image Preprocessing**:
  - `calculate_mean_std`: Calculates dynamic mean and standard deviation for per-channel normalization.
  - `load_image_as_tensor`: Loads images from paths, resizes them, and normalizes based on calculated mean and standard deviation.
- **Model Initialization**:
  - `CLIPModel` and `CLIPTokenizer` encode text and visual context.
  - `StableDiffusionPipeline` leverages the denoising diffusion process for image generation.
- **ContextDiffusion Class**: Combines text and visual context through encoding and concatenation, guiding the diffusion model for image generation.

Results

In-Domain Tasks

The model shows high fidelity to visual context when generating images for in-domain tasks. By incorporating multiple visual context images, the model captures intricate details, such as color and style, in the generated image. Context Diffusion significantly outperforms traditional models in FID (Fréchet Inception Distance) and RMSE (Root Mean Square Error) scores, especially in scenarios where text prompts are not provided.

Out-of-Domain Tasks

In out-of-domain tasks, such as converting sketches to images, Context Diffusion effectively transfers style and details from context images. Human evaluations indicated a preference for images generated by Context Diffusion, with a win rate of over 55% compared to competing models.

Figure 1: In-domain comparison to Prompt Diffusion [48]: Examples of HED, seg mentation, depth-to-image as forward tasks and image-to-HED, segmentation, depth as reverse tasks, with both visual context and prompt given as conditioning information.

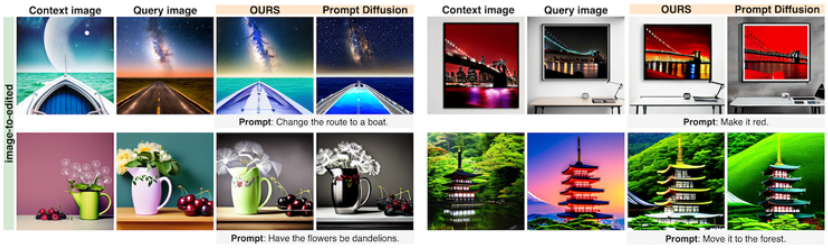

Figure 2: Out-of-domain comparison to Prompt Diffusion [48]: sketch, normal map, scribble, canny edge-to-image tasks. Visual context and prompt (C+P) are given as conditioning information.

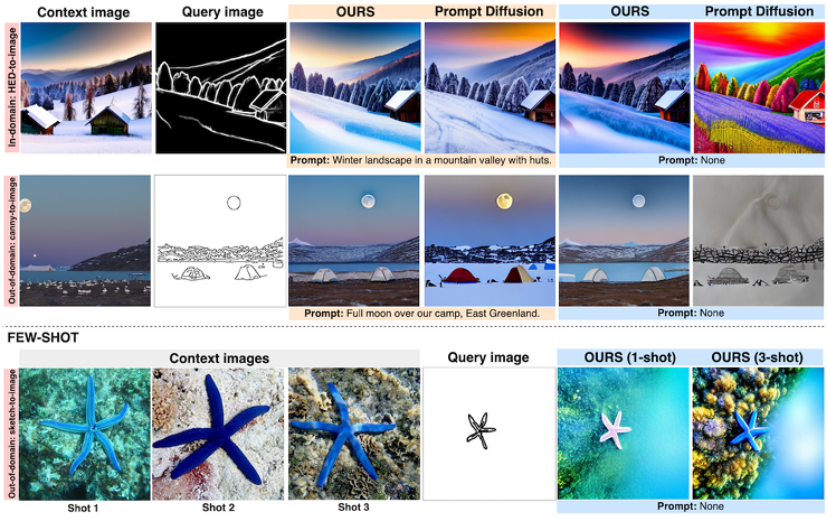

Figure 3: General

Mathematical Justification

The mathematical strength of Context Diffusion lies in its cross-attention mechanism, which enables the model to leverage both text and visual context by learning a combined embedding:

$$z_t = z_t + \text{CrossAtt}(Q = z_t, K = V = [h^c, h^V]) \tag{3}$$

where:

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

By concatenating text and visual embeddings, the model can manage single or multiple context images, providing high-quality outputs even without text prompts. This setup is key to the model's adaptability across various tasks.

Conclusion

Context Diffusion presents a robust framework for generating images directly from visual context, with optional text prompts. By integrating multiple visual context images and a query layout, the model learns strong contextual cues that ensure high fidelity in the generated images. Its adaptability makes it suitable for applications like media content creation and custom image generation.

Significance

Context Diffusion broadens the capabilities of in-context learning in image generation by enabling models to focus on visual context alone. This approach has significant implications in areas where text prompts are unavailable, allowing for the creation of contextually rich images based purely on visual examples.

Future Directions

Future work could expand Context Diffusion by exploring:

- **Real-time Control Mechanisms**: Enabling dynamic visual editing during image generation.
- **Higher Resolution Training**: Training on higher resolution images to improve detail and contextual precision.
- **Extended Few-Shot Settings**: Experimenting with additional context images for even greater flexibility in generation.

Output

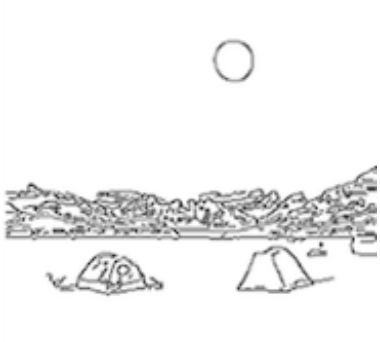

Figure 4: Query Image

Figure 5: Context Image

Figure 6: Output Image

prompt = "Full moon over our camp, East Greenland"