CSYE: 7380

Latent Diffusion Models

VAE, UNet and Fine-tuning CLIP

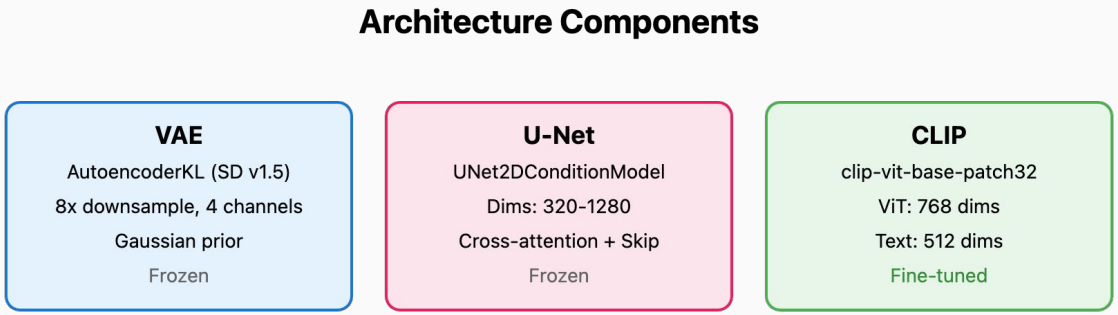
Shashwat Shahi
shahi.sh@northeastern.edu

Architecture Components

1. **Variational Autoencoder (VAE)**
- AutoencoderKL (SD v1.5)
- 8x downsample, 4 channels
- Gaussian prior
- Frozen

2. **U-Net Model**
- UNet2DCondition (SD v1.5)
- Latent diffusion
- Cross-attention, skip connections
- Frozen

3. **CLIP Model**
- clip-vit-base-patch32
- Vision: 768 dims, 32x32 patches
- Text: 512 dims, 12 layers
- Fine-tuned on Flickr8k

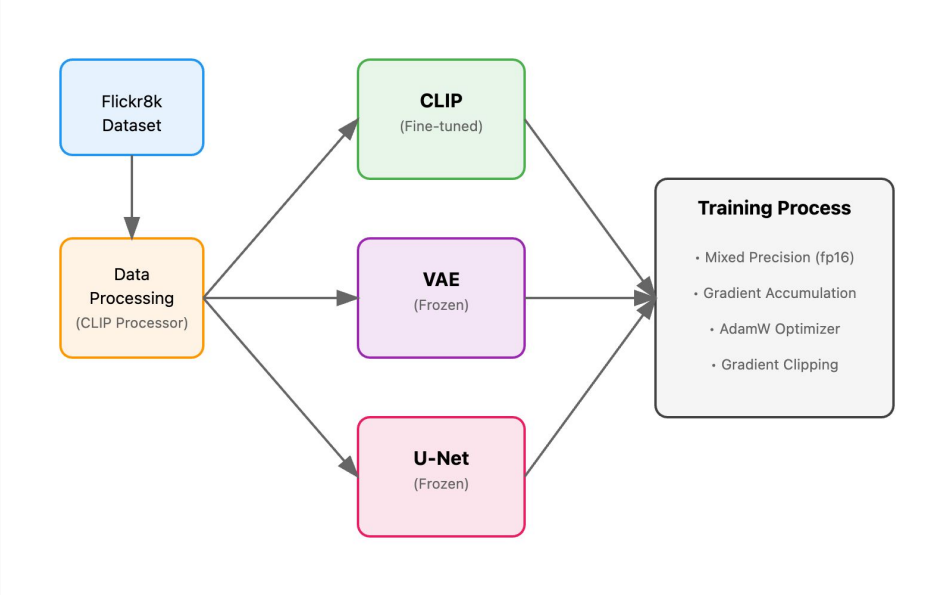


Implementation Details: Training Pipeline

1. Data Preparation
- Load and preprocess images
- Process text captions
- Create batches with proper padding
- Apply CLIP-specific preprocessing

2. Training Loop
- Mixed precision training (fp16)
- Gradient accumulation
- Regular checkpointing

3. Optimization
- AdamW optimizer
- Learning rate: 1e-5
- Gradient clipping at 1.0
- Batch size: 32



Implementation Details: Fine-Tuning Strategy

1. **Model Selection:** CLIP undergoes fine-tuning while VAE and U-Net stay frozen.

2. **Direct Data Path:** Flickr8k dataset flows through CLIP processor before entering models.

3. **Optimized Training:** Uses fp16 precision, gradient accumulation, and AdamW optimizer.

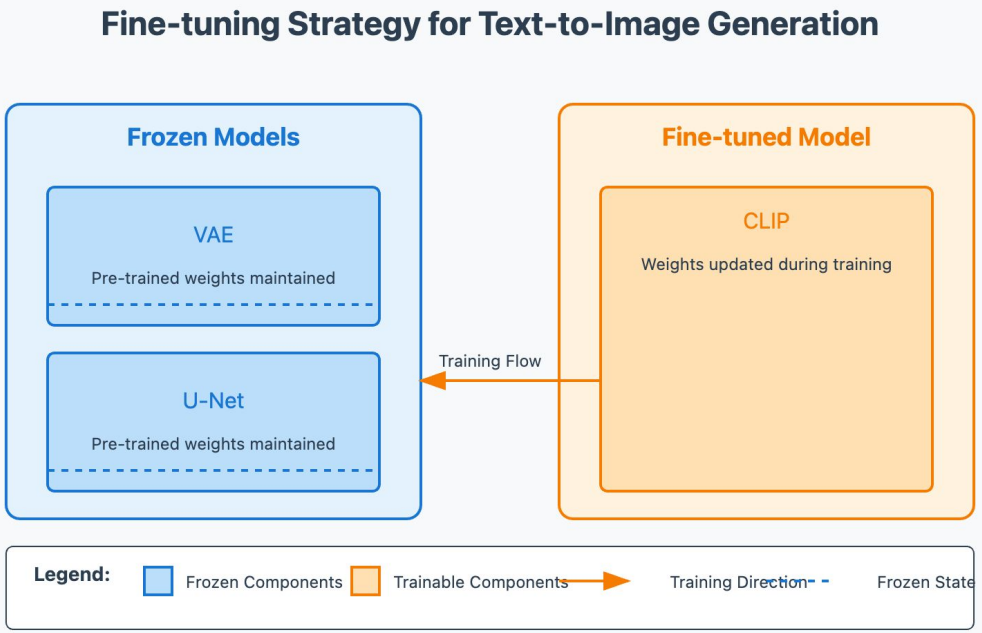
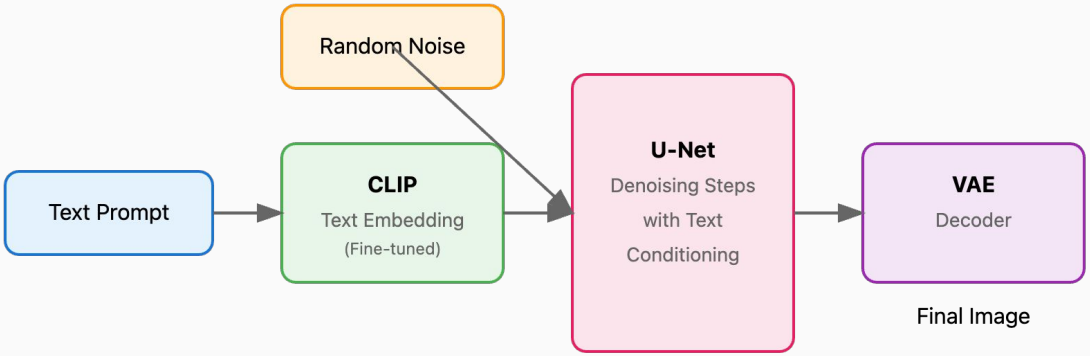


Image Generation Process

1. **Input Processing:** Text prompt is converted to embeddings by CLIP while random noise is generated as a starting point for the image creation.

2. **Guided Denoising:** U-Net gradually refines the random noise into meaningful content, using the text embeddings as guidance to ensure the image matches the description.

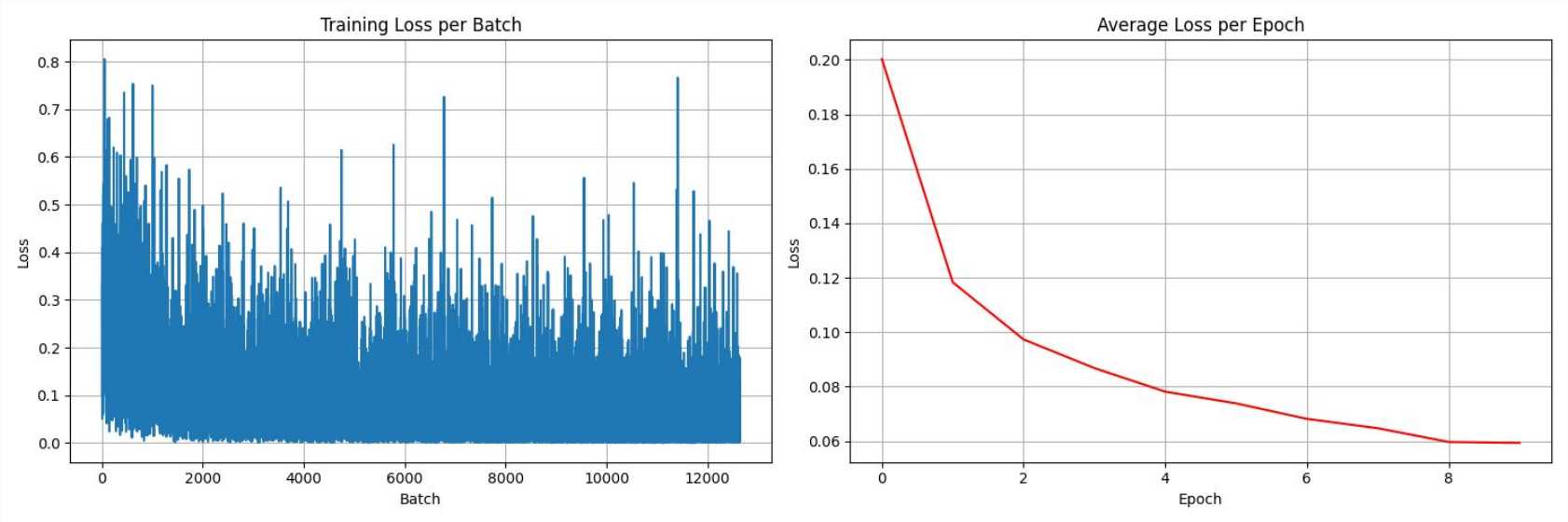
3. **Final Rendering:** VAE decoder transforms the refined latent representation into a high-quality, full-resolution image that aligns with the original text prompt.



Results and Visualizations: Training Loss

- The training graphs demonstrate the model's learning progression, where the batch-wise training shows fluctuating but generally decreasing loss (blue graph).

- The epoch-level average loss (red graph) reveals a smooth, consistent decline in error rate from 0.20 to approximately 0.06 over 9 epochs, indicating successful model convergence.



Results and Visualizations: Prompt Results

Prompt: A dog running on the beach at sunset

Prompt: A colorful garden with blooming flowers

Thank You