Imagen: Text-to-Image Diffusion Model

Mustafa Hajij

June 20, 2025

What is Imagen?

- A generative model for creating high-quality images from text prompts.
- Models the conditional distribution $p(x|t)$, where:
  - x: Image ($H \times W \times C$).
  - t: Text prompt.
- Uses a diffusion process to transform noise into images.

Goal
Photorealistic images aligned with text descriptions.

Forward Diffusion Process

- **Purpose**: Gradually add noise to an image x_0.
- Over $T \approx 1000$ steps:

$$x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

- $\beta_t \in (0, 1)$: Cosine-based noise schedule in Imagen.
- Result: $x_T \approx \mathcal{N}(0, I)$ (pure noise).

Reverse Diffusion Process

- **Purpose**: Reconstruct image from noise x_T.
- Learn denoising distribution:

$$x_{t-1} \sim p_\theta(x_{t-1}|x_t, t) = \mathcal{N}(\mu_\theta(x_t, t), \beta_t I)$$

- Neural network predicts noise: $\epsilon_\theta(x_t, t, c)$.
- Guided by text embedding c.

Noise Schedule: Cosine-Based

- **What is it?**: The noise schedule controls how much noise is added at each step.
- **Cosine Schedule**: Used in Imagen for smoother transitions.
- **Why it matters**: Affects image quality and training stability.
- **Formula**: $\beta_t = 1 - \frac{\cos(\alpha_t)}{\cos(\alpha_{t-1})}$, where α_t is a cosine function.

Impact
Cosine schedule leads to better sample quality than linear schedules.

T5: Text Encoder

- **T5 (Text-to-Text Transfer Transformer)**: A pre-trained transformer model.
- **What it does**: Converts text prompts into high-dimensional embeddings.
- **Embeddings**: Capture the meaning of the text to guide image generation.
- **Why it matters**: Ensures the generated image matches the text description.

Key Point
T5 provides the ”instructions” for the diffusion model.

Text Conditioning

- **Text Encoder**: T5-XXL (11B parameters) maps text t to embedding $c = f(t)$.
- **Conditioning**: Denoising network uses c:

$$\epsilon_\theta(x_t, t, c)$$

- Implemented via cross-attention in U-Net.

U-Net Architecture

- **Backbone**: U-shaped CNN with encoder-decoder and skip connections.
- **Components**:
  - Residual blocks for local features.
  - Attention layers for global context.
  - Cross-attention for text embeddings c.
- **Scale**: Millions of parameters for high-quality output.
- **Time Encoding**: Sinusoidal embeddings for timestep t.

Cascaded Diffusion: Step-by-Step Resolution

- **Goal**: Generate high-resolution images (1024x1024) efficiently from text.
- **Process**:
  - **Base Model**: Generates a 64x64 image from noise, conditioned on text embeddings.
  - **First Super-Resolution U-Net**: Upscales 64x64 to 256x256, adding medium-level details.
  - **Second Super-Resolution U-Net**: Upscales 256x256 to 1024x1024, refining to high fidelity.
- **Why multiple U-Nets?**:
  - Direct 1024x1024 generation is computationally expensive and error-prone.
  - Progressive upscaling splits the task:
    - Base model captures structure.
    - First U-Net enhances clarity.
    - Second U-Net polishes details.
  - Improves efficiency and image quality.

Example
Text: ”A golden retriever in a beret.”

Training Dataset

- **LAION-400M**: A large-scale, open-source dataset of image-text pairs.
- **Proprietary Data**: Additional image-text pairs for enhanced diversity.
- **Importance**: Provides the variety needed for the model to generalize across different prompts.

Key Point
The quality and size of the dataset directly impact the model’s ability to generate diverse, high-quality images.

Training Imagen

- **Objective**: Train each U-Net to predict noise for its specific resolution step (64x64, 256x256, 1024x1024).
- **Compute**: Hundreds of TPUs for large-scale training.
- **Noise Schedule**: Cosine schedule enhances sample quality.

Improving Image Fidelity

- **Classifier-Free Guidance**:

$$\hat{\epsilon}_\theta = \epsilon_\theta(x_t, t, \emptyset) + s \cdot (\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \emptyset))$$

- Guidance scale $s \in [3, 10]$.
- **Dynamic Thresholding**: Adjusts pixel values to avoid saturation.

Outcome
Superior FID and text-image alignment vs. DALL-E 2.

Generating Images

1. Start with noise $x_T \sim \mathcal{N}(0, I)$.
2. For $t = T$ to 1:
   - Predict noise $\hat{\epsilon}_\theta(x_t, t, c)$.
   - Sample $x_{t-1} \sim \mathcal{N}(\mu_\theta(x_t, t, c), \beta_t I)$.
3. Upsample via cascade: 64x64 → 256x256 → 1024x1024.

Result
Photorealistic, text-aligned images.

Applications and Limitations

- **Applications**:
  - Creative design (e.g., art, advertising).
  - Data augmentation for machine learning.
  - Visual storytelling and content creation.
- **Limitations**:
  - Computationally intensive, requiring significant resources.
  - May struggle with complex or abstract prompts.
  - Potential for bias based on training data.