Text-to-Image Generation Using Diffusers Library

- Overview of the text-to-image generation process.
- **Diffusion Models**: Used to synthesize high-quality images based on text prompts.
- Common in applications like *DALL-E*, *SDXL*, and *Janus-Pro*.
- Diffusion transformers enhance the aesthetic and accuracy of generated images.

Diffuser Model Architecture

- Models like *Janus-Pro* use a unified approach for multimodal understanding and visual generation.
- Independent encoding is applied for tasks like visual generation and language instruction.
- Visual decoder converts image tokens from the raw input data into coherent images.
- Architecture features include the use of the *VQ Tokenizer* to generate image embeddings and feed them into a large language model.

Optimized Training Process

- **Stage I**: Involves training the adapters and image head.
- **Stage II**: Unified pretraining on both multimodal data and text-to-image data.
- **Stage III**: Supervised fine-tuning focuses on optimizing text-to-image generation while improving multimodal understanding capabilities.
- The use of synthetic aesthetic data accelerates convergence and improves image quality.

Text-to-Image Generation Evaluation

- Models like *Janus-Pro-7B* outperform others in benchmarks such as *GenEval* and *DPG-Bench* for instruction-following image generation.
- **Performance Metrics**:
- GenEval: Measures object-focused alignment and compositional abilities.
- DPG-Bench: Assesses complex prompt generation.
- *Janus-Pro* achieves superior aesthetic quality, stable outputs, and coherent image generation even with challenging prompts.