Introduction

Problem: Users want quick understanding of image content without writing code.
Goal: Build an intuitive web app where users can:

- Upload an image
- Auto-generate natural language captions
- Ask free-form questions about the image

Solution: Combine BLIP vision-language models with Gradio web framework.

Methods & Architecture

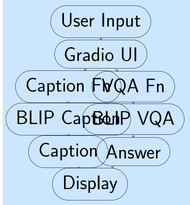
Models Used:

- **BLIP Image Captioning:** `Salesforce/blip-image-captioning-base`
- **BLIP VQA:** `Salesforce/blip-vqa-base`

Framework:

- PyTorch + Hugging Face Transformers
- Gradio (web UI)
- Deployed on Hugging Face Spaces

System Data Flow:



Example Results

Example 1: Street Scene

Caption: "A busy city street with people walking past storefronts."

Question	Answer
How many people?	Several
Time of day?	Daytime

Example 2: Office Setting

Caption: "A workspace with a laptop and coffee mug on a desk."

Question	Answer
Home office?	Yes
Cup color?	White

Assessment: BLIP generates accurate, context-aware captions and answers. Demonstrates understanding of visual content and natural language reasoning.

Key Features

- **Dark-themed UI:** Professional Gradio interface
- **Two-pipeline design:** Separate caption and VQA flows
- **Real-time inference:** Beam search with quality/speed trade-off
- **Easy deployment:** GitHub repo + Hugging Face Spaces
- **Reproducible:** Full source code, requirements.txt, documentation

Implementation

Tech Stack:

Component	Tool
Models	Salesforce BLIP
Deep Learning	PyTorch
Model Hub	HF Transformers
Web UI	Gradio
Deployment	HF Spaces
Language	Python 3.9+

Performance:

- **First run:** ∼1–2 minutes (model download)
- **CPU inference:** 3–5 seconds per query
- **GPU inference:** 0.5–1 second per query
- **Memory:** ∼3.5 GB model weights

Limitations

- Pre-trained models only (no domain fine-tuning)
- Possible hallucination or incorrect answers
- CPU inference is slow; GPU needed for real-time
- Single-image context (no history)
- English-language primary

Future Work

- Domain-specific fine-tuning (medical, retail)
- Multi-turn conversation with memory
- Larger BLIP models or LLM integration
- Evaluation metrics (COCO, human benchmarks)
- Mobile/edge deployment

Conclusion

This project demonstrates how modern multimodal vision-language models can be rapidly integrated into accessible, user-friendly web applications. BLIP models excel at both image understanding and generation, making them ideal for interactive AI tools.

Key Takeaways:

- Pre-trained multimodal models are powerful and accessible
- Simple frameworks (Gradio) enable rapid prototyping
- Real-world applications: accessibility, e-commerce, content understanding