Summary for AI Model Selection

- **BLIP:** High-quality, natural captions with strong Hugging Face and LAVIS support. Supports tone conditioning via text prompts. Ideal balance of quality, speed, and ease of deployment.
- **BLIP-2:** Excellent zero-shot and reasoning ability using large LLM backends, but heavy and requires more GPU memory. Best for advanced multimodal extensions, not MVP.
- **ViT-GPT2:** Lightweight and CPU-friendly. Easy to deploy but produces generic captions. Good fallback or on-device option.
- **OFA/GIT/Florence-2:** Strong captioning and grounding ability but heavier architectures and less straightforward integration for MVPs.

AI Model Selection Reasoning

The selected model will be Salesforce’s BLIP because it offers the best tradeoff between performance, caption quality, licensing, and integration simplicity. The tone control can be implemented through its “prompt conditioning” feature which makes this the more suitable choice.