Model Selection

Chosen Model: Qwen2-VL (Vision-Language Model)

We selected the Qwen2-VL model for our video analytics implementation for several key reasons:

1. **Multimodal Capabilities**: Qwen2-VL excels at processing both visual and textual information, making it ideal for video analysis tasks where understanding both visual content and generating textual descriptions is crucial.
2. **Flexible Model Sizes**: The implementation supports multiple model sizes (2B, 7B, and 72B parameters), allowing for scalability based on computational resources and accuracy requirements:
    - 2B: Lightweight, suitable for rapid prototyping
    - 7B: Balanced performance (our default choice)
    - 72B: Maximum accuracy for critical applications
3. **Zero-shot Performance**: The model demonstrates strong zero-shot capabilities, enabling analysis of diverse video content without specific training for individual use cases.

Implementation Tradeoffs

1. Frame Processing Strategy

Chosen Approach: Interval-based frame sampling

- **Pros**:
    - Reduced computational overhead
    - Faster processing time
    - Lower memory requirements
- **Cons**:
    - Potential to miss rapid events
    - Less granular temporal analysis

Alternative Considered: Processing every frame

- **Pros**: Maximum temporal resolution
- **Cons**: Significantly higher computational cost, redundant analysis of similar frames

2. Analysis Architecture

Chosen Approach: Hybrid analysis system

- Frame-level analysis for temporal queries
- Aggregated summary for general analysis
- **Pros**:
    - Adaptable to different query types
    - Efficient resource utilization
    - Flexible output format

- **Cons**:
    - More complex implementation
    - Potential for inconsistent analysis between modes

3. Hardware Optimization

Chosen Approach: CUDA-aware implementation with CPU fallback

- **Pros**:
    - Optimal performance on GPU-enabled systems
    - Broader compatibility through CPU fallback
    - Automatic device selection
- **Cons**:
    - Additional code complexity
    - Memory management considerations

Technical Implementation Details

1. **Frame Extraction**:
    - OpenCV-based video processing
    - PIL integration for image format compatibility
    - Configurable frame interval (default: 30 frames)
2. **Prompt Processing**:
    - Dynamic prompt type detection
    - Adaptive analysis based on query intent
    - Support for timestamp-specific and general queries
3. **Output Generation**:
    - Structured output format with timestamps
    - Flexible between detailed temporal analysis and summarized results
    - Clean interface for downstream applications