KINETIC: Keyframe Integration for Networked Encoding, Temporal Indexing, and Captioning

Technical Documentation Team
YipYap Project

September 6, 2025

Abstract

We present KINETIC (Keyframe Integration for Networked Encoding, Temporal Indexing, and Captioning), a comprehensive system for integrating video and animated image support into the YipYap media management platform. KINETIC extends the existing image captioning architecture to handle temporal media through intelligent keyframe extraction, temporal caption synchronization, and unified playback controls. Our system introduces novel approaches for frame-accurate caption positioning, multi-modal content analysis across time sequences, and seamless integration with existing bounding box annotation workflows. By leveraging the established caption generation pipeline and extending it with temporal awareness, KINETIC provides a unified experience for managing both static and dynamic media while maintaining the performance characteristics that make YipYap effective for large-scale dataset management.

1 Introduction

Modern media management systems must contend with an increasingly diverse range of content types, from traditional static images to complex temporal media including videos and animated images. The YipYap platform, originally designed for static image management with sophisticated captioning and annotation capabilities, requires significant architectural extensions to handle these temporal media types effectively. KINETIC addresses this challenge by introducing a comprehensive temporal media management system that seamlessly integrates with existing workflows while providing novel capabilities for time-aware content analysis and annotation.

The integration of video and animated image support presents several unique challenges that KINETIC systematically addresses. First, temporal media requires fundamentally different data structures and processing pipelines compared to static images. Second, captioning systems must evolve from single-frame analysis to multi-frame temporal analysis with proper synchronization. Third,

user interface components must provide intuitive controls for navigating temporal content while maintaining the responsive performance characteristics of the existing system. Finally, the system must gracefully handle the increased computational and storage requirements associated with video processing.

2 System Architecture

2.1 Core Components

KINETIC extends the YipYap architecture through five primary subsystems:

1. **Temporal Media Processor**: Handles video decoding, keyframe extraction, and format conversion
2. **Keyframe Management System**: Manages extracted keyframes and their metadata
3. **Temporal Caption Synchronizer**: Coordinates caption generation and positioning across time
4. **Unified Playback Controller**: Provides consistent playback controls across media types
5. **Performance Optimization Layer**: Ensures responsive performance for large video files

2.2 Media Type Extensions

KINETIC extends the existing `ImageModel` to support temporal media through a new `TemporalMediaModel` that inherits from the base model while adding temporal-specific properties:

```
class TemporalMediaModel(ImageModel):
    type: Literal["video"] | Literal["animated_image"]
    duration: float # Total duration in seconds
    frame_rate: float # Frames per second
    keyframes: List[KeyframeInfo] # Extracted keyframes
    temporal_captions: Dict[float, List[Tuple[str, str]]] # Time-indexed
        captions
    playback_metadata: PlaybackMetadata # Encoding, codec info
```

2.3 Keyframe Extraction Strategy

The keyframe extraction system employs a multi-stage approach to identify representative frames for captioning and annotation:

Algorithm 1 Adaptive Keyframe Extraction

```
function ExtractKeyframes(video_path, max_keyframes)
    keyframes ← []
    scene_changes ← DetectSceneChanges(video_path)
    content_changes ← AnalyzeContentVariation(video_path)
    for each frame f in video do
        if IsSceneChange(f) OR IsContentSignificant(f) then
            keyframes.append(CreateKeyframe(f))
        end if
        if len(keyframes) ≥ max_keyframes then
            break
        end if
    end for
    return OptimizeKeyframeDistribution(keyframes)
end function
```

3 Temporal Caption Synchronization

3.1 Multi-Frame Caption Generation

KINETIC extends the existing caption generation pipeline to support temporal analysis through a novel multi-frame approach:

1. **Keyframe Analysis**: Each extracted keyframe undergoes individual caption generation using existing models (JTP2, WDv3, Florence-2)
2. **Temporal Consistency**: Captions are analyzed for consistency across adjacent keyframes to identify stable content descriptions
3. **Time-Based Merging**: Similar captions across time ranges are merged with temporal metadata
4. **Scene Transition Detection**: Caption changes are synchronized with detected scene transitions

3.2 Time-Indexed Caption Storage

The temporal caption system introduces a new data structure for storing time-aware captions:

```
class TemporalCaption:
    start_time: float # Start time in seconds
    end_time: float  # End time in seconds
    caption_type: str # Type of caption (wd, jtp2, etc.)
    content: str     # Caption text
    confidence: float # Confidence score
    keyframe_indices: List[int] # Associated keyframe indices
```

3.3 Playback Synchronization

The playback synchronization system ensures that captions are displayed at the correct temporal positions during video playback:

$$\text{display\_time}(\textit{caption}) = \text{start\_time} + \frac{\text{current\_frame}}{\text{frame\_rate}} \tag{1}$$

4 Unified Playback Controller

4.1 Cross-Media Playback Interface

KINETIC introduces a unified playback controller that provides consistent controls across all temporal media types:

- **Play/Pause Controls**: Standard media playback controls with keyboard shortcuts
- **Frame Navigation**: Frame-by-frame navigation for precise positioning
- **Speed Control**: Adjustable playback speed (0.25x to 4x)
- **Seek Controls**: Click-to-seek and timeline scrubbing
- **Keyframe Jumping**: Quick navigation between extracted keyframes

4.2 Performance Optimizations

The playback system employs several performance optimizations to maintain responsiveness:

1. **Progressive Loading**: Video segments are loaded progressively based on playback position
2. **Memory Management**: Implemented LRU cache for video frames with configurable size limits
3. **Background Processing**: Keyframe extraction and caption generation occur in background threads
4. **Adaptive Quality**: Video quality adjusts based on available bandwidth and system performance

5 Integration with Existing Systems

5.1 Bounding Box Annotation Extension

KINETIC extends the existing bounding box annotation system to support temporal media through frame-accurate positioning:

```
class TemporalBoundingBox(BoundingBox):
    start_frame: int    # Starting frame number
    end_frame: int      # Ending frame number
    keyframe_boxes: Dict[int, BoundingBox] # Boxes per keyframe
    interpolation: InterpolationType # How to interpolate between
        keyframes
```

5.2 Caption Generation Pipeline Integration

The existing caption generation pipeline is extended to support temporal media through a new temporal captioner interface:

```
class TemporalCaptionGenerator(CaptionGenerator):
    async def generate_temporal_captions(
        self,
        video_path: Path,
        keyframes: List[KeyframeInfo]
    ) -> List[TemporalCaption]:
        # Generate captions for each keyframe
        # Analyze temporal consistency
        # Return time-indexed captions
```

5.3 Gallery Integration

The gallery system is extended to support temporal media through new display components:

- **Temporal Thumbnails**: Keyframe-based thumbnails with play indicators
- **Duration Display**: Video duration and frame count information
- **Playback Preview**: Hover-to-play functionality for quick previews
- **Temporal Search**: Search capabilities across temporal content

6 Performance Considerations

6.1 Memory Management

KINETIC implements sophisticated memory management to handle large video files efficiently:

1. **Streaming Decoding**: Video frames are decoded on-demand rather than loading entire files
2. **LRU Frame Cache**: Recently accessed frames are cached with configurable size limits

3. **Background Processing**: Heavy operations like keyframe extraction run in background threads
4. **Memory Monitoring**: Real-time memory usage tracking with automatic cleanup

6.2 Storage Optimization

The system optimizes storage requirements through several strategies:

- **Keyframe Compression**: Extracted keyframes are compressed using WebP format
- **Metadata Caching**: Temporal metadata is cached in SQLite for fast access
- **Incremental Processing**: Large videos are processed incrementally to avoid memory spikes
- **Cleanup Policies**: Automatic cleanup of temporary processing files

7 User Experience Enhancements

7.1 Temporal Navigation

KINETIC provides intuitive temporal navigation through several interface enhancements:

- **Timeline Interface**: Visual timeline showing keyframes and caption positions
- **Keyboard Shortcuts**: Frame-by-frame navigation with arrow keys
- **Seek Preview**: Hover over timeline to see frame previews
- **Bookmark System**: User-defined bookmarks for quick navigation

7.2 Multi-Modal Content Analysis

The system supports multi-modal analysis across temporal content:

1. **Visual Analysis**: Traditional image captioning applied to keyframes
2. **Audio Analysis**: Audio captioning for video content (future extension)
3. **Motion Analysis**: Detection of significant motion patterns
4. **Scene Understanding**: High-level scene classification across time

8 Implementation Details

8.1 Backend Extensions

The backend is extended with new endpoints and data structures:

```
# New API endpoints
@app.post("/api/video/process/{path:path}")
async def process_video(path: str, config: VideoProcessingConfig)

@app.get("/api/video/keyframes/{path:path}")
async def get_keyframes(path: str)

@app.post("/api/video/generate-captions/{path:path}")
async def generate_video_captions(path: str, generator: str)
```

8.2 Frontend Components

New frontend components are introduced for temporal media handling:

- **VideoPlayer**: Custom video player with caption overlay
- **TimelineEditor**: Interactive timeline for caption editing
- **KeyframeGallery**: Grid view of extracted keyframes
- **TemporalBoundingBoxEditor**: Frame-accurate annotation interface

9 Future Extensions

9.1 Audio Captioning

Future versions of KINETIC will include audio captioning capabilities:

- **Speech Recognition**: Automatic transcription of video audio
- **Audio Description**: Generation of descriptive audio for accessibility
- **Multi-Language Support**: Support for multiple audio languages

9.2 Advanced Temporal Analysis

Advanced temporal analysis features planned for future releases:

1. **Action Recognition**: Detection of specific actions and events
2. **Character Tracking**: Consistent character identification across frames
3. **Emotion Analysis**: Temporal emotion tracking
4. **Content Summarization**: Automatic video summarization

10 Conclusion

KINETIC successfully extends the YipYap platform to support temporal media while maintaining the performance and usability characteristics that make the system effective for large-scale dataset management. Through intelligent keyframe extraction, temporal caption synchronization, and unified playback controls, KINETIC provides a seamless experience for managing both static and dynamic media content.

The system's modular architecture ensures that existing workflows remain unaffected while new temporal capabilities are seamlessly integrated. The performance optimizations and memory management strategies ensure that the system can handle large video files efficiently, while the user experience enhancements provide intuitive navigation and annotation capabilities for temporal content.

Future extensions will further enhance the system's capabilities through audio captioning and advanced temporal analysis, positioning KINETIC as a comprehensive solution for multi-modal media management in the YipYap ecosystem.