Multiagent System for Video Evaluation

Technical Report

INTRODUCTION

Recent breakthroughs in generative video synthesis and computer vision have underscored the urgent need for comprehensive and reliable evaluation frameworks capable of assessing video quality across multiple dimensions. In response to this challenge, we introduce a modular **multiagent system** designed to deliver both quantitative metrics and qualitative insights for video evaluation.

Leveraging a distributed, agent-based architecture, the system performs in-depth analysis and produces structured assessment reports—combining automated scoring with interpretive commentary tailored to each specified criterion.

SYSTEM ARCHITECTURE

The Video Content Evaluation System is a comprehensive framework for analyzing video content across multiple dimensions. The system follows a modular, agent-based architecture with publish-subscribe communication:

1. **Core Components**:

   Message Bus: Central communication hub (pub-sub pattern)

   Managing Agent: It manages the whole system based upon the user requirements, it communicates with the user and based upon the user inputs it sends the messages for the respective agents.

   Specialized Agents: Four analysis modules with distinct responsibilities

   Reporting Agent: Gives us the evaluation of both quantitative and qualitative metrics.

Input Video → Frame Extraction → Parallel Analysis → Result Aggregation → Report Generation

2. **Agent Responsibilities**:

   TemporalCoherenceAgent: Frame-to-frame consistency

   SemanticConsistencyAgent: Content-text alignment

   DynamicSceneAgent: Scene transition analysis

   GeneralizationAgent: Content novelty assessment

VIDEO CONTENT EVALUATION ASPECTS:

1. Temporal Coherence (Frame-to-frame Consistency)

Refers to the smoothness and consistency of visual content across consecutive frames in a video. In other words, it ensures that objects, edges, and motions behave naturally over time, without flickering or abrupt changes.

Key Metrics Used:

- **Optical Flow:** Measures the motion magnitude between frames, helping detect abrupt or unnatural movement in generated or real videos.
- **Frame Difference:** Computes the absolute pixel-wise difference between adjacent frames to quantify visual changes.
- **SSIM (Structural Similarity Index):** Evaluates structural similarity between frames by comparing luminance, contrast, and texture patterns.
- **Edge Consistency:** Assesses how well edges align across frames, ensuring continuity of shapes and object boundaries.

2. Semantic Consistency (Content-Text Alignment)

Semantic Consistency refers to the ability of the generated content (like captions or summaries) to maintain meaning and coherence with the input video and its associated reference text. The goal here is to ensure that the summary generated from the video frames aligns well with the reference text, both semantically and contextually.

Key Metrics and models Used:

- **BLIP Model:** BLIP is used for image captioning is used to generate captions for the frames in the video.
- **Sentence Transformer:** Sentence-BERT (SBERT) is used to generate embeddings of the reference text and the generated captions, which are then compared to calculate the semantic similarity.
- **Cosine Similarity:** A common metric to evaluate the semantic closeness between two textual representations (in this case, the reference text and the generated summary).

Purpose: Ensures the video stays true to its intended narrative or description, useful for quality control in automated video generation.

3. Dynamic Scene Analysis (Scene Transition & Motion Handling)

Dynamic Scene Evaluation refers to assessing the movement, changes, and variations in a video to understand the dynamics of the scene. This includes detecting scene changes, analyzing optical flow (camera movement), detecting lighting changes, and tracking object movements to evaluate how well these dynamic aspects are captured and maintained across the video frames.

Key Metrics Used:

- **Scene Change Detection:** It says how often the scene changes based on histogram comparison.
- **Flow Variance:** variation in the optical flow, indicating more or less camera movement.
- **Brightness Changes:** Detects lighting inconsistencies.

- **Object Movement Tracking:** Lucas-Kanade feature tracking.
- **Object movement variance**: variation in object movement between frames

Purpose: Identifies jarring cuts, unstable camera work, or inconsistent lighting that may degrade quality.

4. Generalization (Content Novelty Assessment)

The goal of the **Generalization Agent** is to evaluate how novel the content in a video is by comparing its extracted features to a baseline. This helps identify whether the video content is typical, somewhat novel, or highly novel.

Purpose: Helps determine if the video introduces fresh perspectives or follows predictable patterns, useful for creative evaluation.

Model/Work Process Overview

The evaluation process begins by running the code, which prompts the user to select desired temporal coherence metrics via comma-separated inputs. Next, the Managing Agent requests a reference text describing the video and the video file path. Using these inputs, the system evaluates various aspects of the video, including temporal coherence, semantic alignment, dynamic scene analysis, and content novelty. The results are divided into quantitative metrics (numerical values) and qualitative assessments (descriptive feedback). Finally, the Reporting Agent compiles and saves the evaluation in a JSON report, providing a comprehensive summary of the video's overall quality.

Experimental Result for Evaluating One Video

Video Path:

/WhatsApp Video 2025-05-01 at 15.42.00_19c1ab1c.mp4

Reference Description:

"A person kneeling on the floor lighting a cracker for some time and he went on his bed with his friends."

Selected Metrics:

1. Optical Flow
2. Frame Difference
3. SSIM
4. Edge Consistency

Quantitative Metrics:

- **Optical Flow**: 0.7910
- **Frame Difference**: 6.7818

- **SSIM**: 0.9276
- **Edge Consistency**: 0.2026

Semantic Metrics:

- **Semantic Score**: 0.5682
- **Summary**: "A person kneeling on the floor, a man kneeling on the floor, a woman is standing on the floor in a room, a man laying on a bed with a laptop, a person is standing in a room with a fire, a person is holding a sparkle stick, a person is sitting on the floor with their feet up, a group of people sitting on a couch, a person laying on a bed with a remote control."
- **Reference**: "A person kneeling on the floor lighting a cracker for some time and he went on his bed with his friends."

Dynamic Metrics:

- **Scene Change Ratio**: 0.0854
- **Flow Variance**: 0.4521
- **Brightness Change Frequency**: 0.0339
- **Average Object Movement**: 13.04
- **Object Movement Variance**: 851.5725
- **Assessment**: "Moderate handling - noticeable artifacts"

Generalization Metrics:

- **Novelty Score**: 1.0
- **Assessment**: "Highly novel content"

Qualitative Assessment:

- **Temporal Quality**: Smooth
- **Semantic Alignment**: Moderate alignment
- **Scene Continuity**: Moderate handling - noticeable artifacts
- **Content Familiarity**: Highly novel content

Discussion:

The project successfully integrates multiple evaluation modules to assess video quality across temporal, semantic, dynamic, and generalization dimensions. It provides both quantitative scores and qualitative insights, making the evaluation more comprehensive. The user-friendly interface allows for flexible metric selection, and the final report summarizes the video’s coherence, relevance, and novelty. This system is effective for validating generated or edited videos in research and practical applications