# ChameleonCast

Realistic Video to Cartoon Transformation

Srujan Agrawal (sru28@seas.upenn.edu)

CIS 5810 Fall 2025 Final Project - Track 1 Custom

## 1. Project Title and Summary

ChameleonCast is a video transformation pipeline that converts realistic videos into cartoon representations by replacing human actors with animated animal characters while preserving actions and narrative structure. This project extends the image-to-text-to-image workflow (Track 1.6) into the temporal domain, addressing unique challenges in maintaining spatial consistency and temporal coherence across video frames.

**Project Goal:** The primary application is educational content for children, particularly in sports education. By converting game footage into cartoon format, coaches and parents can make sports concepts more engaging and accessible to young learners. The cartoon style captures children's attention while maintaining the instructional value of the original content.

**Key Achievements:** Successfully developed a frame-based segmentation approach that significantly improved upon baseline results. The final system processes videos through keyframe transformation and multi-image context generation, producing cartoon videos with better spatial accuracy and action fidelity than text-only approaches.

## 2. Pipeline and Baseline Results

### 2.1 Baseline Implementation

The initial baseline pipeline followed a straightforward three-stage architecture:

- **Stage 1: Video Analysis** - Input video processed through Gemini API to generate comprehensive descriptions of actors, actions, settings, and narrative flow.
- **Stage 2: Cartoon Prompt Generation** - Gemini API transformed realistic descriptions into cartoon-style prompts, replacing humans with animated animals while preserving actions.
- **Stage 3: Video Generation** - Text-to-video model (tested with Runway and HeyGen) generated cartoon output from prompts.

### 2.2 Baseline Results

Testing revealed significant limitations with the text-only approach:

- **Spatial Inaccuracy:** Generated videos failed to accurately reproduce specific locations and spatial relationships. Character positions and environmental details diverged significantly from source.
- **Action Loss:** Complex movements and interactions were oversimplified or altered, losing critical details necessary for educational content.
- **Temporal Drift:** Longer videos showed increasing divergence from source material over time, with actions and scenes bearing little resemblance to the original by the end.
- **Inconsistent Styling:** Character designs and animation styles varied throughout videos, creating jarring visual discontinuities.

These baseline results demonstrated that text prompts alone provide insufficient guidance for maintaining fidelity to source videos. Visual anchors were clearly necessary for acceptable results.

# 3. Improved Approach: Frame-Based Segmentation

## 3.1 Methodology

To address baseline limitations, I developed a frame-based approach leveraging both visual and textual guidance:

- **Video Segmentation:** Divided input videos into 4-second segments (Veo 3.1 maximum duration) to maintain temporal coherence.
- **Keyframe Extraction:** Extracted first and last frames from each segment as visual anchors.
- **Image-to-Image Transformation:** Converted first frames to cartoon style using image-to-image models, establishing consistent visual style while preserving spatial details.
- **Multi-Image Context:** Utilized Gemini 2.5's multi-image capability to analyze both the cartoon-styled first frame (style reference) and original last frame (location/action reference) simultaneously, generating accurate segment descriptions.
- **Guided Video Generation:** Fed cartoon keyframes and enhanced prompts to Veo 3, which supports beginning and ending frame guidance along with text prompts.
- **Segment Stitching:** Combined transformed segments into final output video.

## 3.2 Technical Rationale

This approach addresses each baseline limitation systematically:

- **Spatial Accuracy:** Keyframes provide explicit visual anchors that preserve locations and compositions from source video.
- **Style Consistency:** Image-to-image transformation on first frames ensures uniform cartoon aesthetic across all segments.

- **Action Preservation:** Multi-image analysis allows the system to maintain both desired style and specific actions/movements from source material.
- **Temporal Coherence:** Short segments (4 seconds) with frame guidance produce smooth, continuous motion within each segment.

## 4. Experimental Results and Analysis

### 4.1 Experiment 1: First Frame Transformation

**Objective:** Establish reliable cartoon style conversion for segment openings while preserving spatial details.

**Method:** Applied image-to-image transformation to first frames of video segments. Tested multiple animation style prompts to find optimal balance between cartoon aesthetic and detail preservation.

**Results:** Successfully achieved high-quality cartoon conversion with excellent detail retention. The animated frames captured character positions, environmental elements, and action states accurately. This established a consistent visual style that served as the foundation for subsequent video generation.

**Key Finding:** Image-to-image models proved far superior to text descriptions for maintaining spatial fidelity while achieving stylistic transformation.

### 4.2 Experiment 2: Last Frame with Multi-Image Context

**Objective:** Generate accurate ending frames that maintain both style consistency (from first frame) and location fidelity (from source video).

**Method:** Used Gemini 2.5's multi-image processing capability. Provided the model with: (1) cartoon-styled first frame as style reference, and (2) original last frame for location and action details. The model generated descriptions incorporating both inputs.

**Results:** Achieved good quality last frame transformations with notable improvement in spatial accuracy compared to single-image approaches. The dual-reference method effectively balanced style transfer with location preservation, though some minor discrepancies remained in complex scenes.

**Key Finding:** Multi-image context significantly improved the model's ability to maintain spatial relationships while applying consistent styling.

### 4.3 Experiment 3: Video Generation with Veo 3

**Objective:** Evaluate whether frame guidance combined with text prompts produces higher-quality video than text-only baseline.

**Method:** Generated video segments using Veo 3 with: (1) cartoon-styled first frame, (2) cartoon-styled last frame, and (3) text description of actions. Compared results against baseline text-only generation.

**Results:** Frame-guided generation showed improvement over baseline in specific areas—primarily static environmental elements and object locations. Backgrounds and settings maintained better consistency with the source video. However, the critical aspect of character pose and movement preservation remained problematic.

While keyframes provided anchors at segment boundaries, the models struggled to interpolate accurate motion between them. Dynamic actions, complex movements, and character interactions still diverged significantly from the source material. The improvements were insufficient for the intended use case of educational content where action fidelity is essential.

### 4.4 Experiment 4: Pose Estimation Approach (Unsuccessful)

**Objective:** Attempt to extract precise actor positions and movements using pose estimation, then feed this structured data to the video generation model for more accurate action reproduction.

**Method:** Explored using pose estimation models via the Replicate API to track skeletal positions across frames. The plan was to convert these pose coordinates into detailed text descriptions or conditioning inputs that would guide the video generation model to match exact movements.

**Results:** This approach proved impractical for several reasons:

- **API Reliability Issues:** The Replicate API showed inconsistent performance with frequent timeouts, variable latency, and occasional failures that made it unsuitable for a reliable pipeline.
- **Pose Data Translation:** Even when pose estimation worked, translating skeletal coordinate data into actionable guidance for video generation models was extremely challenging. The models couldn't effectively consume raw pose coordinates.
- **Limited Applicability:** Pose estimation only worked well for human figures in clear view. It struggled with occlusions, multiple actors, and non-human elements, making it unreliable for diverse content.
- **Complexity vs. Benefit:** The additional pipeline complexity and processing overhead didn't justify the marginal improvements in accuracy compared to the simpler frame-based approach.

**Conclusion:** This experiment demonstrated that direct visual keyframe guidance (frames themselves) provides better results than attempting to extract and re-encode positional information through pose estimation. The frame-based approach proved both simpler and more effective.

### 4.5 Experiment 5: Segment Length Optimization

**Objective:** Determine optimal segment duration for balancing quality and processing efficiency.

**Method:** Tested segment lengths from 2 to 6 seconds. Initial experiments used longer segments; final implementation adopted 4-second segments per Veo 3.1 specifications.

**Results:** 4-second segments provided optimal balance. Shorter segments (2s) required excessive stitching and increased processing time without significant quality gains. Longer segments (5-6s) showed temporal drift and action inconsistencies. The 4-second duration maintained coherent motion while keeping action sequences tightly coupled to keyframes.

### 4.6 Challenge: Gemini API Limitations

**Issue:** The Gemini API, while powerful for multi-image analysis and prompt generation, introduced significant practical challenges that impacted development and testing.

**Cost Implications:** Gemini API costs accumulated quickly, especially when processing multiple frames per video segment. Each segment required at least two API calls (analyzing original video content and generating cartoon prompts from multi-image context), and longer videos with more segments multiplied these costs substantially. This made extensive experimentation and iteration expensive, limiting the number of test runs feasible within project constraints.

**Rate Limiting:** More critically, Gemini's rate limits caused significant development delays. The API enforced strict request-per-minute quotas that became bottlenecks when processing segmented videos. Processing a single minute of video broken into 4-second segments required 15+ API calls, frequently triggering rate limits. This forced the implementation of wait periods and retry logic, dramatically increasing total processing time and making rapid iteration during development frustrating and slow.

**Impact on Development:** These limitations constrained the experimental approach. Testing variations in prompt engineering, comparing different segmentation strategies, or evaluating quality across multiple videos became time-consuming and costly. The rate limits particularly affected the ability to quickly iterate on the multi-image analysis pipeline, as each failed attempt consumed quota and required waiting before retrying.

**Workarounds:** Implemented caching of Gemini responses where possible and prioritized careful prompt design over volume of experiments. However, the fundamental constraint remained a significant factor in development velocity.

## 4.7 Challenge: Video Stitching

**Issue:** While individual segments showed good quality, combining them into seamless final videos proved challenging. Transitions between segments occasionally showed minor visual discontinuities.

**Approach:** Implemented frame blending at segment boundaries and ensured last frame of segment N matched first frame of segment N+1 in style and composition. This reduced but did not eliminate all transition artifacts.

**Outcome:** Achieved acceptable transition quality for most content. Complex scenes with rapid movement remain challenging, representing an area for future improvement.

## 5. Project Timeline and Milestones

The project was completed over a 10-week period, progressing through systematic development and refinement phases:

| Week | Phase | Completed Work |
|---|---|---|
| 1-2 | Baseline Development | • Implemented text-to-video baseline pipeline<br>• Integrated Gemini API for video analysis<br>• Set up Runway/HeyGen testing environment<br>• Documented baseline limitations |
| 3-4 | Frame-Based Approach | • Developed video segmentation algorithm<br>• Implemented keyframe extraction<br>• Tested image-to-image transformation models<br>• Validated first frame conversion quality |
| 5-6 | Multi-Image Integration | • Integrated Gemini 2.5 multi-image processing<br>• Developed last frame transformation pipeline<br>• Conducted Veo 3 frame-guided generation tests<br>• Implemented 4-second segmentation |
| 7-8 | Optimization & Refinement | • Optimized segment length (4-second)<br>• Refined prompt engineering strategies<br>• Developed segment stitching algorithm<br>• Conducted comparative quality analysis |
| 9-10 | Testing & Documentation | • Completed full pipeline testing<br>• Analyzed results and documented findings<br>• Prepared final report and presentation<br>• Identified future improvement directions |

### Key Milestones Achieved

| Milestone | Completion | Deliverable |
|---|---|---|
| Baseline Pipeline | Week 2 | Functional text-to-video system with documented limitations |
| Frame Transformation | Week 4 | Reliable first frame cartoon conversion |
| Multi-Image System | Week 6 | Integrated last frame processing with dual-reference context |
| Complete Pipeline | Week 8 | Full segmentation + transformation + stitching system |
| Final Evaluation | Week 10 | Comprehensive results analysis and documentation |

## Conclusions and Future Directions

**Project Outcomes:** While ChameleonCast successfully demonstrated that frame-based guidance improves certain aspects of video-to-cartoon transformation compared to text-only approaches—particularly in object and location preservation—the project ultimately fell short of achieving reliable 1:1 video transfer. The system showed improvements in spatial accuracy for static elements and environmental details, but the fundamental challenges of video generation and character pose preservation proved more difficult than anticipated.

**What Worked:**

- Frame-based keyframe extraction improved object and location preservation compared to baseline
- Multi-image context with Gemini 2.5 provided better environmental consistency
- Static elements and backgrounds showed notable improvement in cartoon conversion
- Pipeline architecture successfully integrated multiple APIs and processing stages

**What Didn't Work:**

- Character pose and movement preservation degraded significantly across frames, especially in dynamic scenes
- Video generation models struggled to maintain action continuity even with keyframe guidance
- Complex character interactions and multi-actor scenes produced unreliable results
- The gap between keyframes (4 seconds) was still too large for maintaining precise pose continuity
- Overall output quality fell short of being usable for the intended educational application

**Key Contributions:**

- Identified that frame-based guidance improves static element preservation but is insufficient for dynamic action transfer
- Demonstrated the limitations of current text-to-video models for pose-accurate content transformation
- Documented that 4-second segments are too coarse-grained for maintaining character pose continuity
- Established that visual keyframes alone cannot bridge the semantic gap for complex motion
- Provided empirical evidence that the video-to-cartoon problem requires more sophisticated approaches than image-to-image techniques extended temporally

**Fundamental Limitations:**

- Character pose and movement fidelity was the critical failure point—the system could not maintain accurate body positions and motion across frames
- Current text-to-video generation models lack fine-grained control necessary for pose-accurate transformations

- Frame-based guidance proved insufficient—keyframes 4 seconds apart cannot constrain intermediate motion
- Complex multi-actor scenes and dynamic actions produced outputs that diverged significantly from source material
- Gemini API rate limits and costs significantly constrained experimentation and iteration
- Processing time (8-10 minutes per minute of video, plus rate limit delays) made rapid testing impractical

**What Would Actually Be Required:**

- Explicit motion and pose conditioning in video generation models—keyframes alone are insufficient
- Temporal consistency mechanisms that maintain character pose across longer sequences without drift
- Significantly shorter segment duration (potentially frame-by-frame guidance) to constrain motion
- Video diffusion models specifically trained on style transfer tasks with pose preservation
- Alternative approaches entirely—such as pose-guided animation systems or rotoscoping-style techniques
- More sophisticated APIs or local models to eliminate cost and rate limit constraints on experimentation

**Final Assessment:** This project represents an honest exploration of the video-to-cartoon transformation problem and demonstrates that the challenge is significantly more complex than initially anticipated. While the frame-based segmentation approach showed promise for preserving static elements and environmental details, it ultimately failed to achieve the core goal of maintaining character pose and action fidelity across temporal sequences. The project's value lies not in a successful solution, but in systematically identifying why current approaches fall short: text-to-video models lack the fine-grained motion control necessary, keyframe guidance alone cannot bridge large temporal gaps, and the semantic understanding required to translate actions across style domains exceeds current model capabilities. Future work in this space will require fundamentally different approaches—potentially involving explicit motion models, pose-conditioned generation, or video diffusion models with stronger temporal consistency mechanisms.