# Decoding Media for the Next Generation:

A Theory-Based and Machine Learnings Approach to Understanding and Shaping Children's Media

Santiago Won Siu

A Thesis Submitted to the Creative Computing Institute

MSc Data Science & Artificial Intelligence

University of the Arts London

Advisor: Tim J. Smith

September 2024

## Abstract

The increasing influence of media on children's cognitive and emotional development has raised concerns about its impact, making it essential to classify and shape media in ways that support healthy growth. This thesis develops a machine-learning-driven feature extraction model to analyse children's media, providing a new framework for understanding its complexities. The model is designed to classify media based on age-appropriateness and to identify features that contribute to cognitive demands, both positive and negative.

Inspired by theories on children's media and executive function, the thesis hypothesizes that cognitively demanding features such as number of characters and objects, scene transitions, luminance shifts, and continuity in places and objects increase as the target age of media increases. The methodology involves collecting theoretical research on children's media, building a multimodal feature extraction tool to analyse visual content, and employing a logistic regression model to predict age classifications based on extracted features.

The findings are presented with visualizations to show patterns in the data, and the limitations of the model are discussed. In addition to age classification, this thesis explores future possibilities, such as using generative models to modify media based on optimal parameters regarding suitability and complexity. The results advance understanding of children's media, highlighting the importance of identifying both features that promote growth and those that may harm cognitive development.

## Declaration

This thesis is my original work and has not been submitted in any previous application for a degree. All work included in this thesis was carried out by me, except where otherwise stated.

## Table of Contents

# I. Introduction

The increasing role of media in shaping children's cognitive and emotional development has sparked concern and research into the potential benefits and harms of early media exposure. My master's journey began with a personal goal: to create **something impactful and scalable for children**. Early discussions with my thesis advisor, Professor Tim Smith, led us to explore the development of an age classification model for children's media, addressing key concerns like **protecting cognitive processes and attention spans**.

The direction was significantly influenced by Professor Smith's recent collaborative work with Claire Essex, particularly their paper "Understanding the Differential Impact of Children's TV on Executive Functions: A Narrative-Processing Analysis" (Essex, Gliga, Maninda, & Smith, 2022). This research highlights the cognitive harm that exposure to certain media features, such as **rapid scene transitions** and fantastical elements, can cause, especially by **overloading attention** and working memory in younger viewers.

This thesis follows a **hypothesis-driven approach**, predicting that cognitively demanding features—such as **faster scene transitions, more muted colour palettes, and complex narratives—will be more prevalent in media aimed at older children.** In contrast, content for younger children is expected to use **slower pacing, brighter colours, simpler narratives, and exaggerated facial expressions** to aid their cognitive and emotional processing. These predictions are rooted in the research "How Infants Perceive Animated Films" (Kirbas & Smith, 2018), which found that younger audiences respond more to visually simplified and exaggerated content designed to guide immature gaze and comprehension.

However, my personal goals extend beyond shielding children from potentially harmful media. Conversations with researcher Jodie Jackson further reinforced my belief in the **positive potential of media**. I now see media not only as a potential threat to children's development but also as a powerful tool for fostering personal growth. This has shaped my vision for a model that not only identifies harmful elements but could eventually promote content that supports cognitive and emotional development.

In line with this goal, the thesis builds a **feature extraction model that assesses content based on cognitive demands and emotional engagement**, focusing on features like scene transitions, semantic and narrative complexity, and colour distribution. Valkenburg's work on "The Development of a Child into a Consumer" (Valkenburg & Cantor, 2002) also provides critical insight, showing how children's preferences evolve toward more emotionally complex content as they age, reinforcing the hypothesis that older-targeted media will feature more intricate narratives and less visually exaggerated cues.

By developing this model, I aim to create a tool that helps media companies, content creators, parents, and educators better understand and curate content that supports children's growth. The findings from this research will not only advance understanding of media's impact on children's cognitive development but will also pave the way for building tools that can actively promote positive developmental outcomes.

# II. Methodology

In this thesis, my work proceeded through several key stages to achieve the goal of developing a framework and tool for interpreting and classifying children's media.

1. **Building the Theoretical Framework based on Research and Trends**:
   I conducted review of existing research, theories, and expert opinions regarding media for children, in order to understand the current landscape, trends, effects, and risks associated with children's media consumption. This included academic papers, interviews with experts, and relevant case studies.

   Based on this research, I assembled a qualitative framework for age classification, which serves as the foundation for the following stages of my work.

2. **Developing a Feature Extraction Tool**
   I developed a multimodal feature extraction tool that collects visual input to analyse media content. This tool identifies key features that help in interpreting the complexity and appropriateness of media content for children.

3. **Database Gathering**
   I gathered a database of media content that had already been age-classified and feature-classified by experts. This dataset provided a reference to assess the feature extraction tool and a benchmark for evaluating the accuracy of my model.

4. **Data Analysis and Model Development**
   I analysed the extracted data using several methods:
   - I ran data visualizations to illustrate key findings and patterns in the media content.
   - I built a predictive regression model to classify media based on age, measuring the model's accuracy using the training data.
   - I also ran a multiple logistic regression analysis for each feature to understand how each contributes to the overall age classification.

   I then discussed the findings in detail, including the limitations of the current model and feature extraction process.

5. **Tools exploration for Future Work Potential**
   I explored different tools such as SAM 2 and stable diffusion models with the aim of understanding the potential of not only flagging harmful content but being able to correct it.

# III. Trends: How Media Threatens Children's Development

Several major trends are reshaping the media industry, including the rise of new and massive content channels, the exponential increase in content generation, and the incorporation of emerging technologies like generative AI (GenAI) tools. In an environment where regulation struggles to keep pace with innovation, new risks emerge, suggesting that advanced **technologies to regulate and assess content may be more crucial than ever**.

With more platforms delivering media content, children are increasingly exposed to vast amounts of it. For example, YouTube sees over 3.7 million videos uploaded daily, while TikTok exceeds 24 million, contributing to a staggering **10 billion videos generated annually across just these two platforms** (Hayes, 2024) (Ch, 2024)**.**

In parallel, the role of **GenAI** in generating content presents growth (Murphy, 2024) and new challenges. These models often function as **black boxes**, creating content exposed to children without clear insight into the decision-making processes behind it. This raises critical concerns about how such content can be ensured to be safe, appropriate, unbiased, and educational. Research groups, such as Anthropic, are focused on tackling the challenges of interpretability, working on ways to extract interpretable features and better understand their behaviour within neural networks (Templeton & Conerly, 2024).

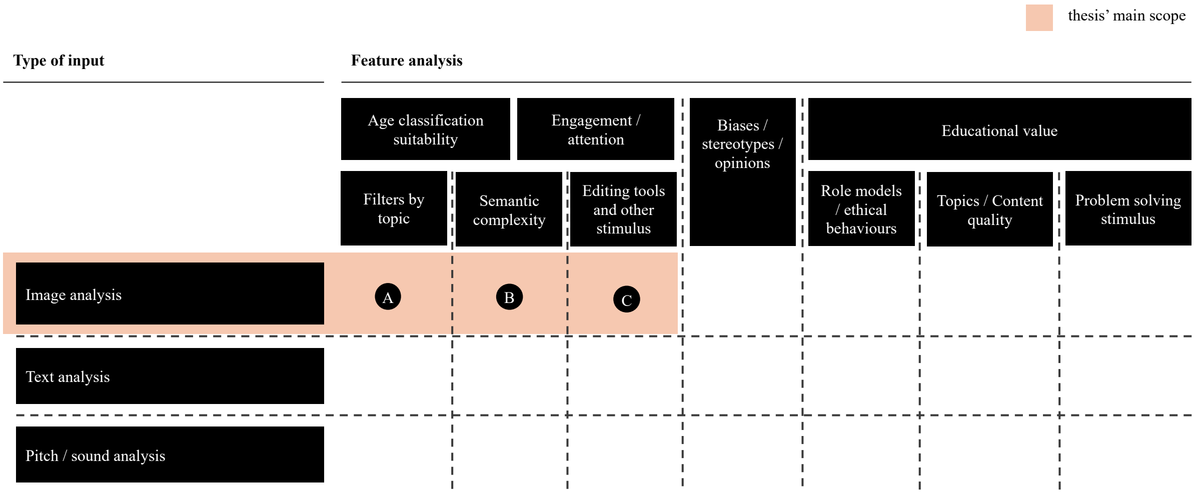
While the journey toward fully interpretable neural networks remains a work in progress, the question arises: what if we could define enough exhaustive feature parameters to minimize all risks? Even without full understanding of how content is generated, we might begin to **"over-parametrize" media**, promoting transparency and potentially transforming black boxes into "glass boxes" (Rai1, 2019).

# IV. Theoretical Framework: Main dimensions for Analysing the Impact of Media on Children

To analyse the impact of media on children, four main dimensions have been identified: age suitability filters and complexity, the effects on attention spans and cognitive processes from editing tools and stimulus design, biases, and educational value.

## Age Classification and Feature Extraction

**Image 1: Matrix to analyse Feature's Impact of Media on Children**



*Source: Self-elaboration*

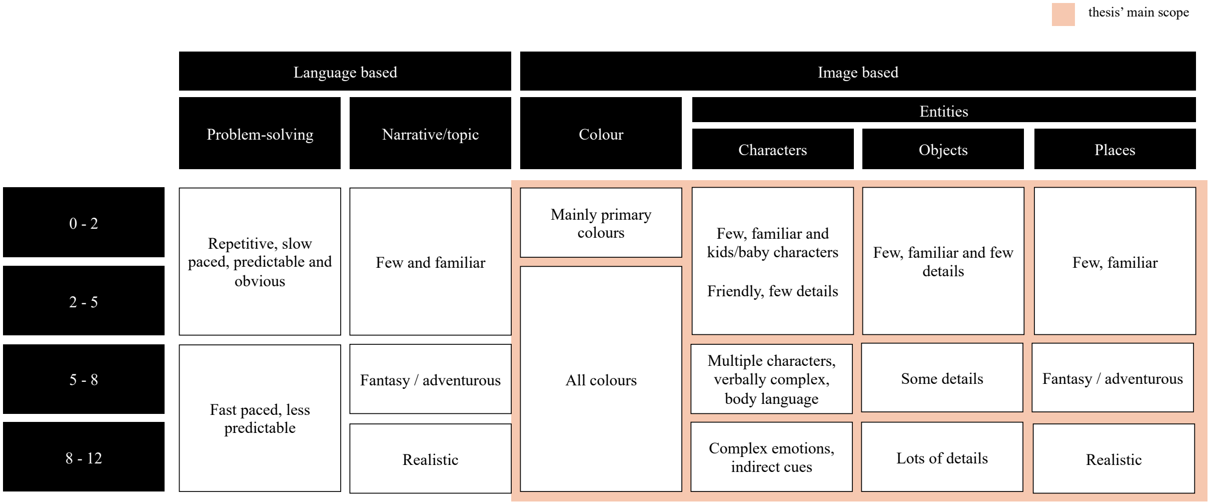
For the purposes of this thesis, the scope focuses on three specific sub-dimensions (labelled A, B, and C in Image 1). These areas support assessments on age classification, semantic complexity, and the proper use of editing tools and stimulus that foster engagement and maintain attention. These three sub-dimensions are crucial in guiding the features extracted from media to make accurate assessments:

1. **Age Classification Filters:** According to experts from CommonSense Media, age classification can be determined by assessing filters (Image 1, Section A) such as profanity, violence, scariness, sex, romance, nudity, drinking, drugs, smoking, and product placements. The CommonSense Media website contains over 1,500 detailed children's animated TV show reviews, each providing both qualitative and quantitative assessments of these features. This dataset serves as an invaluable resource not only for selecting

features to extract but also in case fine-tuning age classification algorithms were required (Common Sense Media, 2024).

2. **Semantic Complexity**: Semantic complexity also plays a crucial role in determining the appropriateness of media for different age groups. By synthesizing existing research, a framework was developed to assess this complexity (Image 1, Section B). This part of the analysis focuses on the visual elements within the media. Key features extracted include colors, the number and characteristics of entities, edge detection, and whether the entities are based in fantasy or reality (Image 2) (Essex, Gliga, Maninda, & Smith, 2022) (Valkenburg & Cantor, 2002).

**Image 2: Semantic Complexity Framework**



▨ thesis’ main scope

| | Language based | | Image based | | | |
|---|---|---|---|---|---|---|
| | | | | Entities | | |
| | Problem-solving | Narrative/topic | Colour | Characters | Objects | Places |
| 0 - 2 | Repetitive, slow paced, predictable and obvious | Few and familiar | Mainly primary colours | Few, familiar and kids/baby characters<br>Friendly, few details | Few, familiar and few details | Few, familiar |
| 2 - 5 | | | All colours | | | |
| 5 - 8 | Fast paced, less predictable | Fantasy / adventurous | | Multiple characters, verbally complex, body language | Some details | Fantasy / adventurous |
| 8 - 12 | | Realistic | | Complex emotions, indirect cues | Lots of details | Realistic |

*Source: Content* (Valkenburg & Cantor, 2002)*]and Self-elaboration*

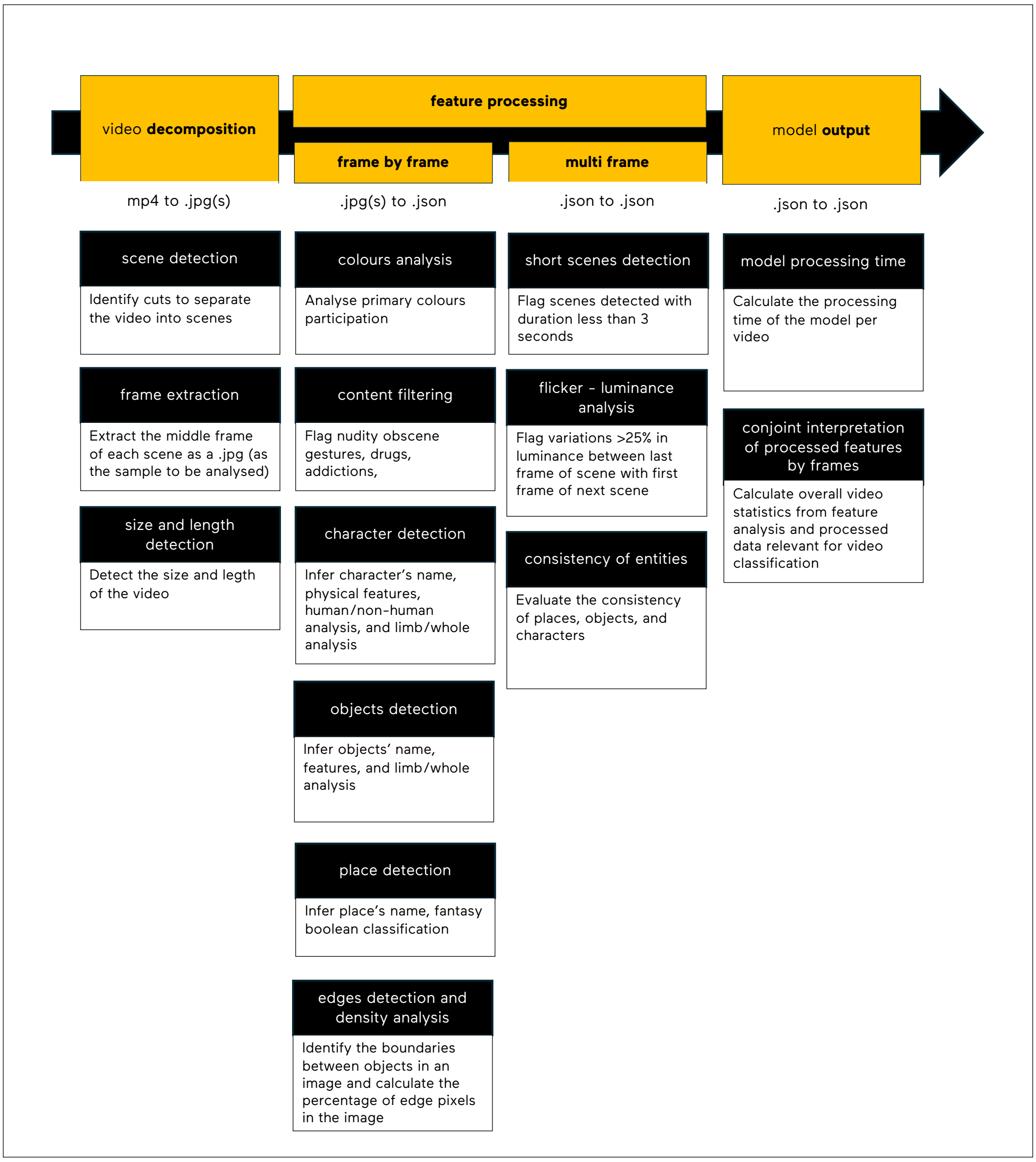
3. **Editing Tools and Stimulus**: Lastly, the impact of editing tools and stimulus on children's engagement and attention spans are explored. The analysis covers features such as flicker and luminance patterns, the frequency of short scenes, and the average scene length. Although further exploration could include a wider range of editing tools (Image 1, Section C), this thesis focuses on these initial parameters for experimentation (Essex, Gliga, Maninda, & Smith, 2022).

# V. Features Extraction Model

Following thorough research to establish the theoretical framework for this thesis, the focus shifted to building a model capable of extracting the desired features efficiently while ensuring both resource efficiency and result accuracy.

It is important to note that this work is currently limited to image processing and analysis. Text, audio, and pitch are not considered at this stage.

**Image 3: Model's process for feature extraction and assessment of media for children**



video **decomposition**

mp4 to .jpg(s)

scene detection

Identify cuts to separate the video into scenes

frame extraction

Extract the middle frame of each scene as a .jpg (as the sample to be analysed)

size and length detection

Detect the size and legth of the video

**feature processing**

**frame by frame**

.jpg(s) to .json

colours analysis

Analyse primary colours participation

content filtering

Flag nudity obscene gestures, drugs, addictions,

character detection

Infer character's name, physical features, human/non-human analysis, and limb/whole analysis

objects detection

Infer objects' name, features, and limb/whole analysis

place detection

Infer place's name, fantasy boolean classification

edges detection and density analysis

Identify the boundaries between objects in an image and calculate the percentage of edge pixels in the image

**multi frame**

.json to .json

short scenes detection

Flag scenes detected with duration less than 3 seconds

flicker - luminance analysis

Flag variations >25% in luminance between last frame of scene with first frame of next scene

consistency of entities

Evaluate the consistency of places, objects, and characters

model **output**

.json to .json

model processing time

Calculate the processing time of the model per video

conjoint interpretation of processed features by frames

Calculate overall video statistics from feature analysis and processed data relevant for video classification

*Source: Self-elaboration*

The following section explains the sub-processes and techniques to approach each of the desired assessments and outputs.

## 1. Video decomposition

This pipeline detects scenes in children's media and extracts key frames for analysis.

- *analyze_video*: Detects scene boundaries with scenedetect, outputting a list of start and end frames.
- *extract_frames_imageio*: Uses imageio to extract and save the middle frame of each scene for content analysis.
- *get_video_length*: Uses OpenCV to calculate total video duration from frame count and FPS.

This process supports detailed video content analysis to assess its cognitive impact on children. It is relevant to mention that **one of the main suppositions of these whole model** is that the **extracted middle frame** will be maximizing representation of the whole scene while being efficient for image processing in our model.

## 2. Feature Processing (frame-by-frame and multi-frame)

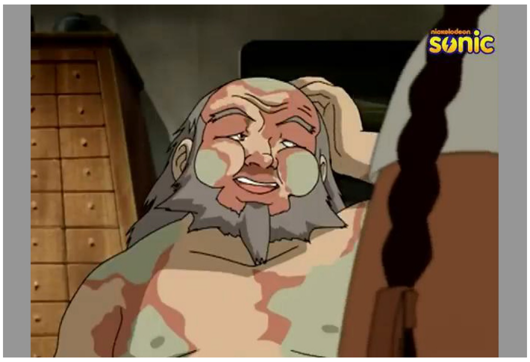
After extracting middle frames, each of them was processed through a list of pre-selected features with the hypothesis that these would be the most efficient regarding assessing the impact of media on children and age classification.

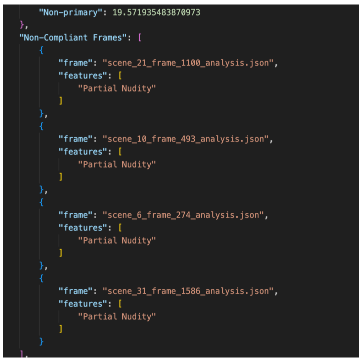
### 2.1. Content Filtering

The model uses OpenAI API's prompt engineering as the technique for processing images and **detecting compliance with filters**. This would ask to look for nudity, obscene gestures, alcohol, drugs, and any sort of addictions.

As a result, filters are flagged frame by frame and then compiled in case there is a non-compliance detection. In the example below, partial nudity is detected.

## Image 4: Example 1 on Filters detection



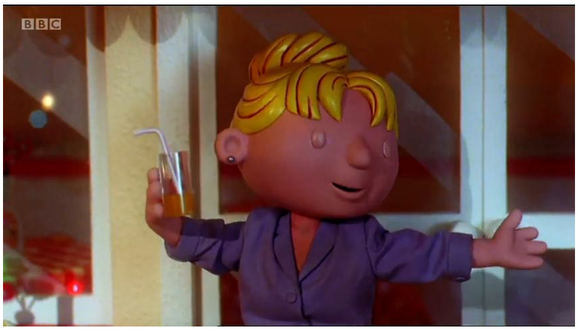


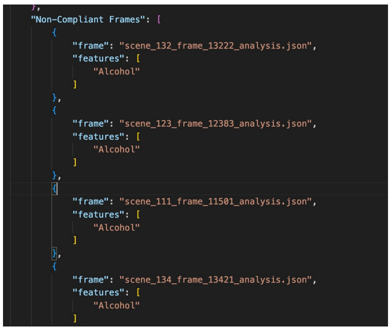
```
        "Non-primary": 19.571935483870973
    },
    "Non-Compliant Frames": [
        {
            "frame": "scene_21_frame_1100_analysis.json",
            "features": [
                "Partial Nudity"
            ]
        },
        {
            "frame": "scene_10_frame_493_analysis.json",
            "features": [
                "Partial Nudity"
            ]
        },
        {
            "frame": "scene_6_frame_274_analysis.json",
            "features": [
                "Partial Nudity"
            ]
        },
        {
            "frame": "scene_31_frame_1586_analysis.json",
            "features": [
                "Partial Nudity"
            ]
        }
    ],
```

*Source: Avatar The Last Airbender*

The **limitation of the model's flagging system** arises when context and text are required alongside image processing. For instance, in an episode of Bob the Builder, the model flags "alcohol" due to a character giving a toast. However, the show doesn't explicitly state whether the beverage is alcoholic or not, highlighting the **model's difficulty in distinguishing contextually nuanced situations**.

## Image 5: Example 2 on Filters detection





```
    },
    "Non-Compliant Frames": [
        {
            "frame": "scene_132_frame_13222_analysis.json",
            "features": [
                "Alcohol"
            ]
        },
        {
            "frame": "scene_123_frame_12383_analysis.json",
            "features": [
                "Alcohol"
            ]
        },
        {
            "frame": "scene_111_frame_11501_analysis.json",
            "features": [
                "Alcohol"
            ]
        },
        {
            "frame": "scene_134_frame_13421_analysis.json",
            "features": [
                "Alcohol"
            ]
```

*Source: Bob the Builder*

## 2.2. Image Complexity and Engagement Assessment

### 2.2.1. Primary Colours Dominance

For assessing the dominance of primary colours, the model analyses each image and returns a structured dictionary that includes the pixel count, percentage of each colour category, and an indication of whether primary or non-primary colours are dominant. The underlying hypothesis is that media targeted at children aged 0–2 predominantly uses primary colours (red, yellow, blue), making this a relevant feature for age classification.

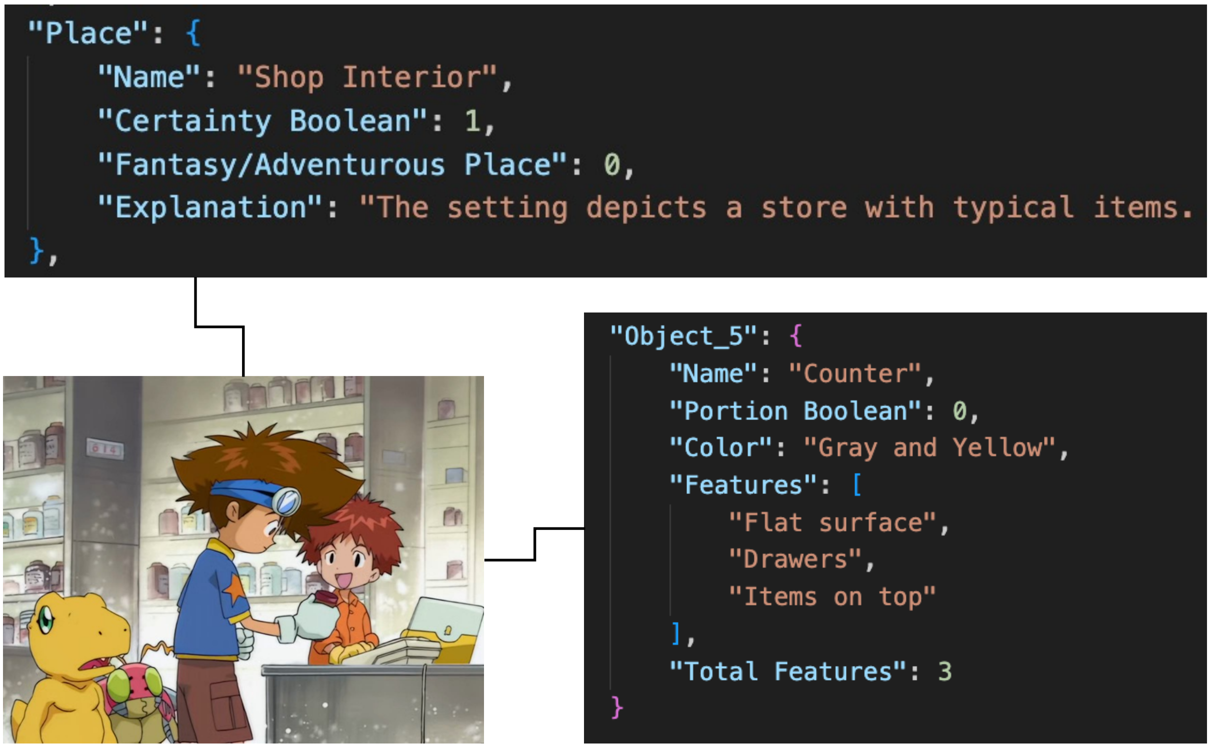
### 2.2.2. Entities Assessment

This thesis refers to **characters, objects, and places as entities**. The selection of assessments is primarily based on Valkenburg's work (Valkenburg & Cantor, 2002), which identifies elements commonly found in media that engage children according to their age. The analysis focuses on the following features:

- The number of entities present during a TV show.
- The complexity of entities (i.e., the amount of detail and features they exhibit).
- The style of entities (i.e., whether they portray fantasy/adventurous or realistic elements).

The latter two features are extracted using prompt engineering on each frame. Once processed, the model calculates an average to provide a final assessment.

**Image 6: Example on Features and Place detection**



```
"Place": {
    "Name": "Shop Interior",
    "Certainty Boolean": 1,
    "Fantasy/Adventurous Place": 0,
    "Explanation": "The setting depicts a store with typical items.
},
```

```
"Object_5": {
    "Name": "Counter",
    "Portion Boolean": 0,
    "Color": "Gray and Yellow",
    "Features": [
        "Flat surface",
        "Drawers",
        "Items on top"
    ],
    "Total Features": 3
}
```

*Source: Digimon Digital Monsters*

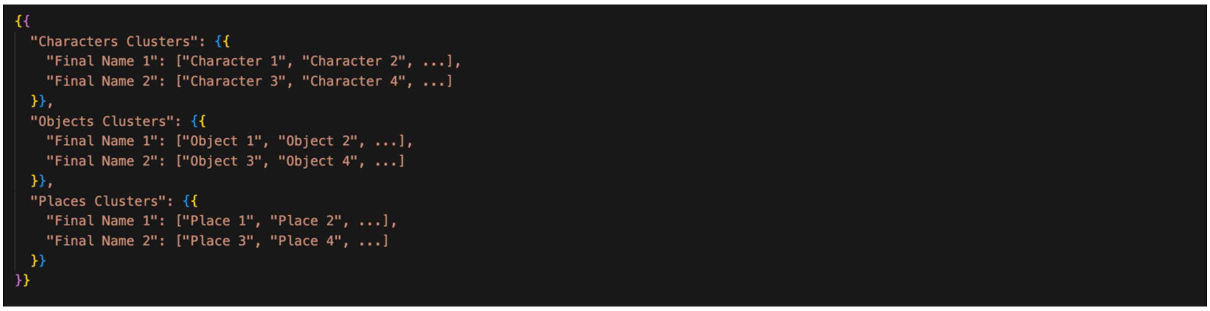
In this case, features were detected for the object "counter." However, some **limitations** are evident:

- "Items on top" are not a physical feature but represent other objects that need to be identified.
- "Drawers" are not visible in the image, which indicates a hallucination by the model.

- The place is classified as non-"fantasy/adventurous", although more context would be required, as the "shop" could exist within an imaginary world. Despite this, 82% of 419 frames from this example were classified as fantasy/adventurous, which aligns with the nature of the show, set in a parallel digital world with fantasy creatures
- Features are also limited by the explicit limit of tokens, thus even if there are more details, the complexity might not be identified properly

To determine the **number of entities in a TV show**, the model had to account for the limitations of the GPT-4 API. The main challenge lies in determining whether an object or character identified across different frames is the same entity.

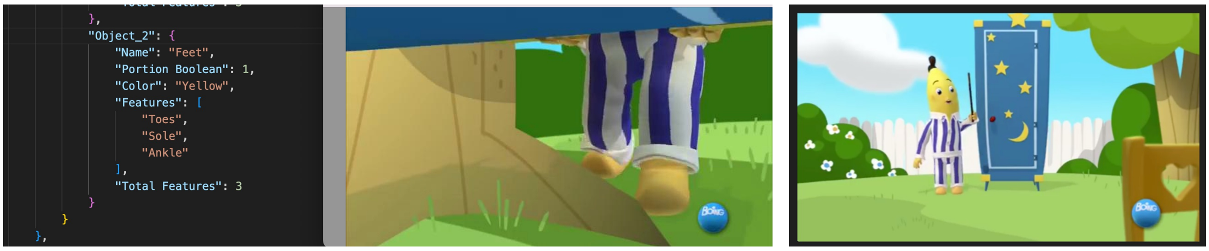
The model processes each frame by prompting the OpenAI API to detect, name, and describe characters and objects. After this, similar but non-identical names might refer to the same entity, so the model consolidates all names into a single file and prompts the OpenAI API to check if any names can be grouped or consolidated, providing a final name.



```
{{
  "Characters Clusters": {{
    "Final Name 1": ["Character 1", "Character 2", ...],
    "Final Name 2": ["Character 3", "Character 4", ...]
  }},
  "Objects Clusters": {{
    "Final Name 1": ["Object 1", "Object 2", ...],
    "Final Name 2": ["Object 3", "Object 4", ...]
  }},
  "Places Clusters": {{
    "Final Name 1": ["Place 1", "Place 2", ...],
    "Final Name 2": ["Place 3", "Place 4", ...]
  }}
}}
```

Additionally, if a frame only displays a part of a character, such as a limb, the model prompts the API with the current frame and frames of other identified entities to verify if the **limb belongs to one of them.**

**Image 7: Example on Limb Detection and Classification**



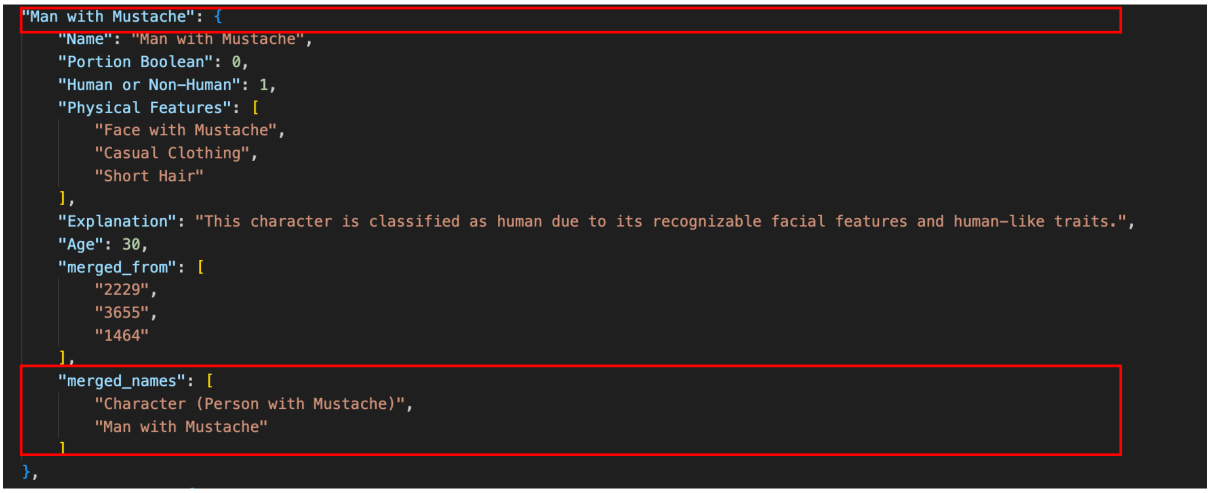
```
            },
            "Object_2": {
                "Name": "Feet",
                "Portion Boolean": 1,
                "Color": "Yellow",
                "Features": [
                    "Toes",
                    "Sole",
                    "Ankle"
                ],
                "Total Features": 3
            }
        }
    },
```

*Source: Bananas in Pyjamas*

One limitation of the model is its approach to object detection. Here, two types of complexity are being assessed:

- **Semantic complexity**: This refers to the number of distinct concepts or objects that the audience must recognise. For example, knowing the general concept of a "car" without needing to differentiate between individual cars is enough for this feature.
- **Narrative complexity**: This refers to tracking characters or special objects within the story, especially following their roles across different scenes. It requires the audience to remember and understand a character's previous role in the narrative.

**Image 8: Example of Clustered Characters**



```
"Man with Mustache": {
    "Name": "Man with Mustache",
    "Portion Boolean": 0,
    "Human or Non-Human": 1,
    "Physical Features": [
        "Face with Mustache",
        "Casual Clothing",
        "Short Hair"
    ],
    "Explanation": "This character is classified as human due to its recognizable facial features and human-like traits.",
    "Age": 30,
    "merged_from": [
        "2229",
        "3655",
        "1464"
    ],
    "merged_names": [
        "Character (Person with Mustache)",
        "Man with Mustache"
    ]
},
```

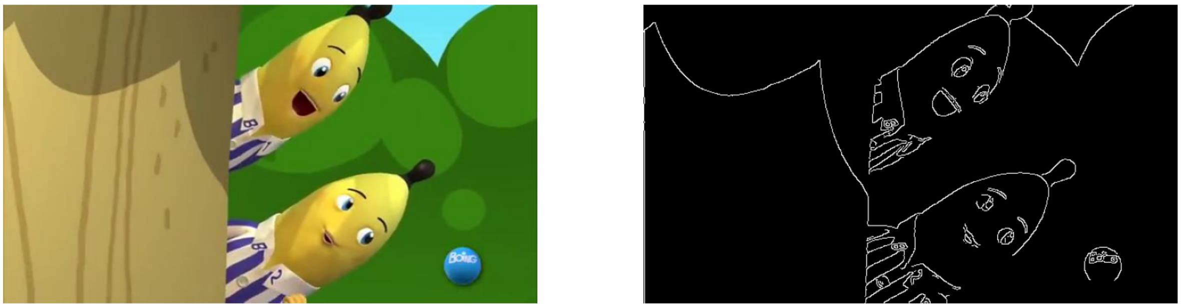
Cluster characters output will keep trace of: previously given names and the frame number per each merged identified character.

While semantic complexity is relatively manageable within the current model, **narrative complexity poses a greater challenge**. The model struggles to differentiate between characters that look alike but are distinct in the story. Addressing this would require an image-by-image analysis of each cropped part of a frame where the character appears. Although computationally feasible, OpenAI's API lacks the precision to accurately crop entities. If this approach were implemented, the model's complexity would increase significantly, from O(n) to O(n$^2$).

### 2.2.3. Edge Density

In addition to the features portrayed in Image2 to assess complexity, the model includes edge density analysis as it is expected to be correlated with the cognitive resources demanded from the viewer (Henderson, Chanceaux, & Smith, 2009).

## Image 9: Example on Edge Detection



*Source: Bananas in Pyjamas*

Previously, this thesis explained how the assessment of entities relies primarily on OpenAI's API. To understand the rationale behind using this API, it is essential to note that several libraries were evaluated to determine their suitability for entity detection.

## Table 1: API's Suitability Assessment for MVP Entities Detection

🟧 thesis' selected API    🟦 Future work potential

| | pros | cons |
|---|---|---|
| **GPT 4o** | • Decent outputs with no objects' pre-training required<br>• Integrated with LLM for interpretation | • Blackbox processing<br>• Inaccuracy because of prompts' reliance<br>• No object segmentation for cropping<br>• Can be expensive for massive processing |
| **OpenCV** | • Flexibility for customized models<br>• Segments the object in the image<br>• Can be fine-tuned to desired needs | • Requires pre-training on extensive animated objects database<br>• Would still rely on LLM integrations for contextual inference / interpretations |
| **Mediapipe** | • Segments the object in the image<br>• Built-in models<br>• Highly efficient | • Low flexibility for customization |
| **AnimeCNN** | • Customized tool for animated video analysis | • Limited to editing tools<br>• Not flexible regarding desired output<br>• No object/entities recognition |
| **TIB AV-Analytics** | • Customized tool for video analysis<br>• Object, places and character recognition | • Not flexible regarding desired output<br>• Limited usage<br>• Low compatibility unless using Ubuntu |
| **SAM 2 by Meta** | • Expected high performance with few or non pre-training | • Would still rely on LLM integrations for contextual inference / interpretations |

*Source: Self-elaborated*

After evaluating various options, the **GPT-4 API was identified as the most suitable for detecting and interpreting entities**, given the limited time and data available to train a custom model. Consequently, prompt engineering has become the primary technique for this task.

However, it is acknowledged that other options, hold greater potential for improving accuracy in future work, particularly due to their flexibility and customisation capabilities. For instance, SAM 2 by Meta presents an interesting approach through image segmentation. SAM 2 can handle sequences of frames, allowing it to segment objects across multiple frames in a video. This makes it suitable for tasks that require consistent object tracking and segmentation throughout a video (Meta AI, 2024).

#### 2.2.4. Entities continuity

This feature calculates the frequency at which characters, objects, and places change between scenes. Frequent changes in these entities require more cognitive effort from viewers to follow the content, indicating higher semantic and narrative complexity. Thus, a greater number of changes between scenes correlates with a TV show's increased complexity.

### 2.3. Image Editing Tools Usage Appropriateness

The model extracts low-level visual features that have been shown to be the strongest exogenous drivers of attention in a scene (Itti, 2007) (Mital, Smith, Hill, & Henderson, 2010). However, this analysis is not intended to establish acceptable or unacceptable thresholds for the impact of these features on children's attention.

#### 2.3.1. Short Scenes Detection:

The model flags scenes with a duration of less than three seconds.

#### 2.3.2. Flicker and Luminance Analysis:

The model uses OpenCV to calculatea the average luminance (brightness) by determining the mean pixel intensity. It then compares the luminance of the last frame of a scene with the first frame of the subsequent scene. If the luminance change exceeds a set threshold (e.g., 25), the transition is flagged as a "strong luminance change" and considered significant. While this threshold is currently in use, further **research is needed to refine its appropriateness**.

**Image 10: Example on Flicker and Luminance Analysis (last vs first frame)**

*Source: Courage the Cowardly Dog*

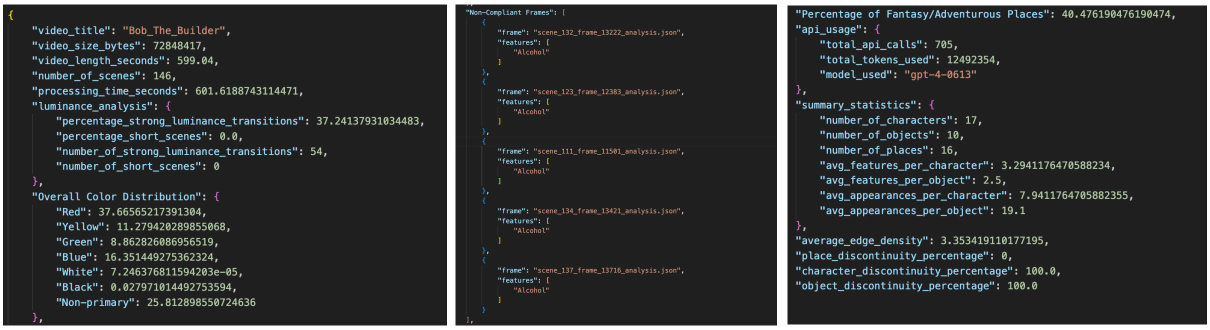
The output of the editing tools assessment will point out how much of these tools is used on average per minute. Which eventually could be used to flag in case there is a suggestion of misuse.

## 3. Model’s Final Output Content

Based on all the previous extracted features and assessments a final .json file will be generated per chapter of TV Show, showing different metrics on its performance.

The main purpose of this data is for it useful to eventually run models on top of it to **classify its appropriateness regarding filters, age classification, and usage of editing tools for the viewer’s visual resources requirement**.

**Image 11: Example of Final Output Data**



```
{
    "video_title": "Bob_The_Builder",
    "video_size_bytes": 72848417,
    "video_length_seconds": 599.04,
    "number_of_scenes": 146,
    "processing_time_seconds": 601.6188743114471,
    "luminance_analysis": {
        "percentage_strong_luminance_transitions": 37.24137931034483,
        "percentage_short_scenes": 0.0,
        "number_of_strong_luminance_transitions": 54,
        "number_of_short_scenes": 0
    },
    "Overall Color Distribution": {
        "Red": 37.66565217391304,
        "Yellow": 11.279420289855068,
        "Green": 8.862826086956519,
        "Blue": 16.351449275362324,
        "White": 7.246376811594203e-05,
        "Black": 0.027971014492753594,
        "Non-primary": 25.812898550724636
    },
```

```
"Non-Compliant Frames": [
    {
        "frame": "scene_132_frame_13222_analysis.json",
        "features": [
            "Alcohol"
        ]
    },
    {
        "frame": "scene_123_frame_12383_analysis.json",
        "features": [
            "Alcohol"
        ]
    },
    {
        "frame": "scene_111_frame_11501_analysis.json",
        "features": [
            "Alcohol"
        ]
    },
    {
        "frame": "scene_134_frame_13421_analysis.json",
        "features": [
            "Alcohol"
        ]
    },
    {
        "frame": "scene_137_frame_13716_analysis.json",
        "features": [
            "Alcohol"
        ]
    }
],
```

```
"Percentage of Fantasy/Adventurous Places": 40.476190476190474,
"api_usage": {
    "total_api_calls": 705,
    "total_tokens_used": 12492354,
    "model_used": "gpt-4-0613"
},
"summary_statistics": {
    "number_of_characters": 17,
    "number_of_objects": 10,
    "number_of_places": 16,
    "avg_features_per_character": 3.2941176470588234,
    "avg_features_per_object": 2.5,
    "avg_appearances_per_character": 7.9411764705882355,
    "avg_appearances_per_object": 19.1
},
"average_edge_density": 3.353419110177195,
"place_discontinuity_percentage": 0,
"character_discontinuity_percentage": 100.0,
"object_discontinuity_percentage": 100.0
```

Furthermore, the output data will also contain **efficiency metrics** such as the amount of **time** to process the whole video file and the amount of **API calls and tokens used**.

# VI. Analysis of Model’s Output

## 1. Summary of the current model’s limitations

- **Limited to Image Processing**: The model only focuses on image processing and analysis, without considering text, audio, or pitch.
- **Middle Frame Assumption**: The model assumes that the middle frame of each scene is representative of the whole scene, which may not always be accurate.
- **Contextual Understanding**: The model struggles with context-sensitive detection, such as distinguishing between alcoholic and non-alcoholic beverages in scenes without textual or contextual clues.
- **Hallucination in Object Detection**: The model sometimes detects features that are not visible in the scene, such as detecting "drawers" that aren't actually present in an image.
- **Lack of Fantasy/Realism Context**: The model flags entities as "fantasy/adventurous" or "realistic" based on visual features, but it lacks the contextual understanding required to differentiate between fantasy and real-world settings, especially in cases where a fantasy setting might visually resemble reality.
- **Semantic vs. Narrative Complexity**: The model can handle semantic complexity (recognising general concepts like a "car") but struggles with narrative complexity, such as tracking a character’s role throughout a story.
- **Inaccurate Cropping:** OpenAI’s API lacks the precision to crop images accurately, making it challenging to perform detailed image-by-image analysis for characters or objects. Thus, it is inaccurate in distinguishing between similar-looking characters
- **Thresholds for Edge Density and Flicker and Luminance**: The current thresholds are provisional and may not accurately reflect what would be significant for children's media consumption, requiring further refinement.
- **Efficiency of Tokens Usage**: The current model does not differentiate between the number of input vs output tokens used for cost estimation. Also it is not registering them appropriately.
- **Objects feature complexity**: The current model is limited by tokens when identifying features, thus it might not be as accurate as required to determine features complexity.

## 2. Efficiency Analysis

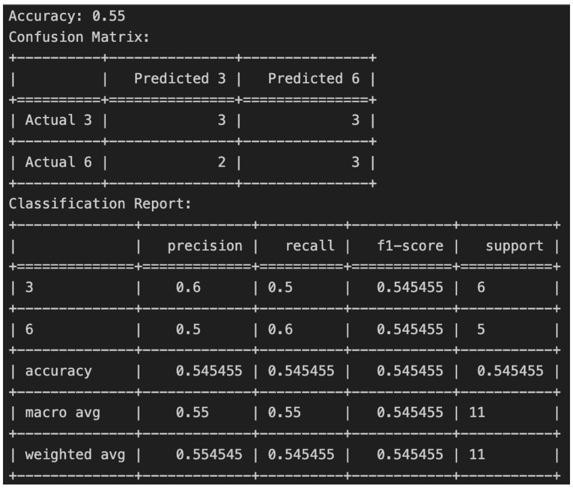
- Average Video Size (Bytes): 216,411,852 bytes
- Average Number of Scenes: 329 scenes
- Average Size per Frame (Bytes): 3,351,091 bytes/frame
- Average cost to process a video: batch 1 - US$2.75 and batch 2 - US$0.23
- Average token used per video:
    - Input: 1,500
    - Output: 1,060

The efficiency analysis highlights the resource demands of video processing. With an average video size of 216,411,852 bytes and 329 scenes per video, the data volume is significant. Each frame averages 3,351,091 bytes, and processing costs average US$2.75 per video, or US$0.008 per frame. These metrics offer a clear **baseline for assessing scalability and cost-efficiency**, pointing to potential optimizations in data handling and processing costs.

It is relevant to analyse that for a second batch of processing, when limiting the max number of analysed scenes/frames per TV Show to 50 for efficiency purposes, the **costs were reduced in 12x to US$0.23 per video**. These are some interesting outputs to eventually look for the minimum required scenes analysis to do age classification. Nonetheless, that wouldn't be exhaustive for filters control.

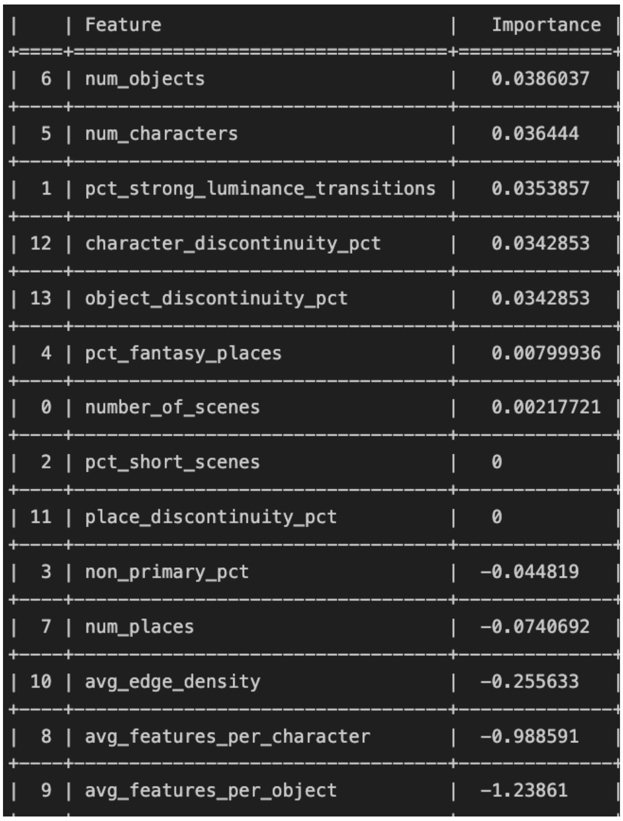
# VII. Analysis and Hypotheses on Age Classification Using Logistic Regression

In this preliminary analysis, logistic regression was applied to **predict age classifications for children's TV shows** using a dataset comprising 30 TV shows for age 3 and 30 TV shows for age 6. The model achieved an **accuracy of 55%,** indicating that while some patterns may be emerging, the current dataset and **features may not be sufficient to reliably distinguish between certain age groups**. The confusion matrix reveals that the model struggles to differentiate between shows aimed at 3-year-olds and 6-year-olds, likely due to similarities in content and themes. It might be relevant to further analyse if these two groups lack relevant enough differences in complexity, themes, and visual elements.



```
Accuracy: 0.55
Confusion Matrix:
+----------+---------------+---------------+
|          |   Predicted 3 |   Predicted 6 |
+==========+===============+===============+
| Actual 3 |             3 |             3 |
+----------+---------------+---------------+
| Actual 6 |             2 |             3 |
+----------+---------------+---------------+
Classification Report:
+--------------+-------------+----------+-------------+-----------+
|              |   precision |   recall |    f1-score |   support |
+==============+=============+==========+=============+===========+
| 3            |    0.6      | 0.5      |    0.545455 |  6        |
+--------------+-------------+----------+-------------+-----------+
| 6            |    0.5      | 0.6      |    0.545455 |  5        |
+--------------+-------------+----------+-------------+-----------+
| accuracy     |    0.545455 | 0.545455 |    0.545455 |  0.545455 |
+--------------+-------------+----------+-------------+-----------+
| macro avg    |    0.55     | 0.55     |    0.545455 | 11        |
+--------------+-------------+----------+-------------+-----------+
| weighted avg |    0.554545 | 0.545455 |    0.545455 | 11        |
+--------------+-------------+----------+-------------+-----------+
```

The feature importance analysis identified **"number of objects" and "number of characters" as relatively significant in influencing the model's predictions**, whereas other features, such as "place discontinuity" and "short scenes", had minimal or no impact. This suggests that while some aspects of content complexity, such as character and object presence, play a role in distinguishing age groups, additional features may be necessary to improve model performance.



```
|    | Feature                            |   Importance |
+====+====================================+==============+
|  6 | num_objects                        |   0.0386037  |
+----+------------------------------------+--------------+
|  5 | num_characters                     |   0.036444   |
+----+------------------------------------+--------------+
|  1 | pct_strong_luminance_transitions   |   0.0353857  |
+----+------------------------------------+--------------+
| 12 | character_discontinuity_pct        |   0.0342853  |
+----+------------------------------------+--------------+
| 13 | object_discontinuity_pct           |   0.0342853  |
+----+------------------------------------+--------------+
|  4 | pct_fantasy_places                 |   0.00799936 |
+----+------------------------------------+--------------+
|  0 | number_of_scenes                   |   0.00217721 |
+----+------------------------------------+--------------+
|  2 | pct_short_scenes                   |   0          |
+----+------------------------------------+--------------+
| 11 | place_discontinuity_pct            |   0          |
+----+------------------------------------+--------------+
|  3 | non_primary_pct                    |  -0.044819   |
+----+------------------------------------+--------------+
|  7 | num_places                         |  -0.0740692  |
+----+------------------------------------+--------------+
| 10 | avg_edge_density                   |  -0.255633   |
+----+------------------------------------+--------------+
|  8 | avg_features_per_character         |  -0.988591   |
+----+------------------------------------+--------------+
|  9 | avg_features_per_object            |  -1.23861    |
```

Given the limitations of the current dataset and features, the **hypothesis** arises that a **larger dataset**, incorporating more **varied features—such as textual or dialogue elements**—might enhance the model's ability to differentiate between age groups. It is also possible that shows targeted at age groups like 3 and 6 are too similar in their structure and content to be easily distinguished by the current feature set, while more pronounced differences exist for broader age group **comparisons, such as 3 and 10**.

Further enhancement can be achieved by incorporating the following features:

- Facial Expressions & Emotional Complexity: Analysis of exaggerated or simplified facial expressions could provide insights into how emotions are conveyed and perceived across different age groups (Kirbas & Smith, 2018)
- Narrative Complexity: Investigating the complexity of narratives, including episodic structure and event continuity, could help assess cognitive demands placed on different age groups (Essex, Gliga, Maninda, & Smith, 2022).
- Character Dynamics: Studying the emotional range and interactions between characters would offer a deeper understanding of how media fosters emotional and social development (Essex, Gliga, Maninda, & Smith, 2022)
- Animation Realism: Future work could compare life-like versus cartoonish animations to determine their impact on recognition and cognitive load (Kirbas & Smith, 2018)
- Consumerism in Media: Analysing the presence of consumerist messages could reveal how children's attitudes towards materialism are shaped by media (Valkenburg & Cantor, 2002)

Incorporating these features would further improve the model's ability to analyse children's media from cognitive and social perspectives.

# VIII. Future Work Routes

## Increasing model's accuracy and scope of assessment

Future work for this thesis will broaden the current assessments of media by focusing on two main areas:

**Model Accuracy:**

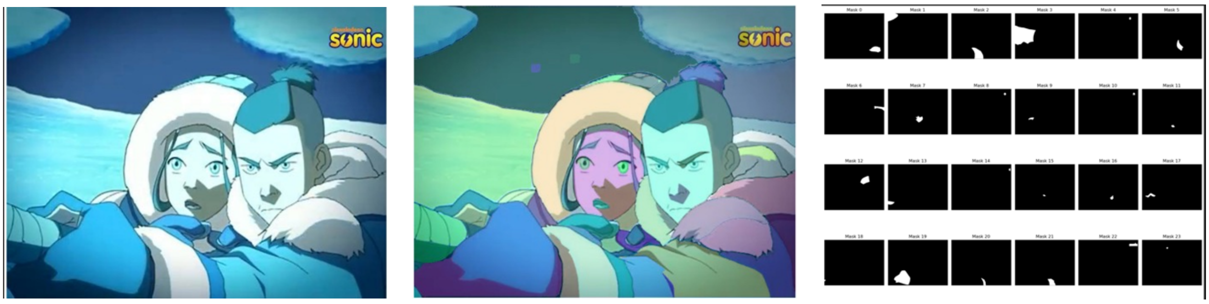
- Incorporating diverse data sources, such as dialogue interpretation through speech diarization, speech-to-text technologies, and pitch tone analysis.

- Integrating more image features related to complexity, including the recognition of emotions and facial gestures.
- Fine-tuning prompts to improve overall accuracy.
- Investigating the use of SAM 2 by Meta for consistent object segmentation across multiple frames, which may serve as a partial replacement for gpt4o by OpenAI.

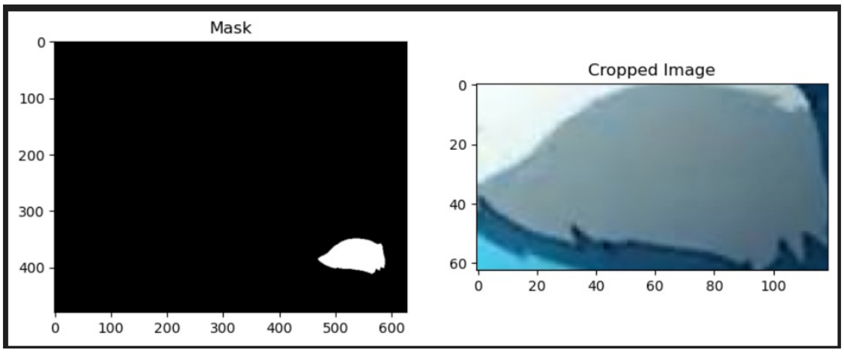
**Scope of Assessments:**

- Evaluating the presence of biases and educational content within media, aligned with the explored theoretical framework.

Work with the SAM model has been, which can be used to extract masks of any frame, that combined with other GenAI model such as Gpt4o, show a strong potential.



*Source: Avatar The Last Airbender*

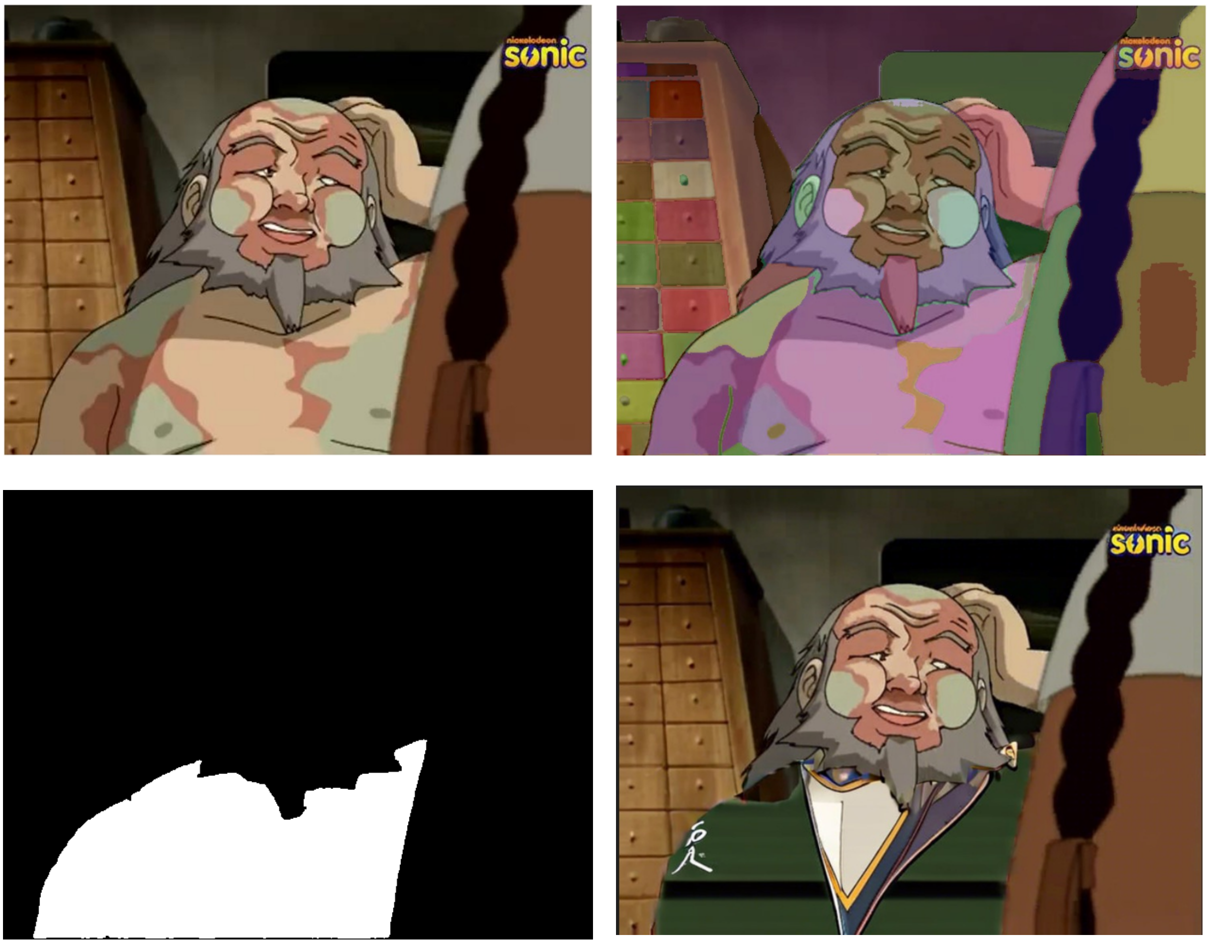


The current limitations of this thesis' model stem from its heavy reliance on the precision of entity and feature identification by gpt4o. By breaking down these components and prompting comparisons with the entire frame, we can expect to achieve greater precision in the analysis of future work.

## Generating “Corrected” Media

The next step involves exploring the potential of combining the SAM model's segmentation capabilities with the interpretability that gpt4o can derive from these segments, alongside the inpainting features offered by Stability AI through Stable Diffusion models.

The following example demonstrates how this multimodal tool identifies nudity in the TV Show “Avatar: The Last Airbender” and generates new content to address it.



This combination could revolutionize content creation by enabling seamless adjustments while preserving artistic integrity. Additionally, it opens up opportunities for enhancing storytelling and adapting media for audiences with disabilities, such as image interpretation for visually impaired children, making content more accessible and inclusive.

# IX. Conclusions

This thesis highlights the complex relationship between media exposure and children's cognitive and emotional development. While certain media features can be detrimental, the potential for media to foster growth is significant. This model opens the path for a transparent tool to understand the impact of a wide range of media features.

The current model's limitations, particularly in contextual understanding and multi-modal analysis, emphasize the need for ongoing refinement. Expanding the dataset and exploring qualitative differences in media content will be crucial for developing a more nuanced framework for age classification.

Moreover, integrating generative media techniques through the SAM model and Stable Diffusion for corrections opens new avenues for content creation that can positively impact children’s development. This research serves as a foundational step toward creating effective tools that shape a media landscape supporting children’s growth while balancing protective measures against harmful content.

Finally, these feature extraction tool and classification model have practical implications for media companies, educators, and parents, allowing for more informed content curation that aligns with developmental needs. Ethical considerations in media production and regulation are vital, ensuring content remains engaging yet developmentally appropriate.

# X. References

Ch, D. (2024). *TikTok Statistics: Revenue & Usage* . Retrieved from https://sendshort.ai/statistics/tiktok/

*Common Sense Media*. (2024). Retrieved from https://www.commonsensemedia.org/tv-reviews

Essex, C., Gliga, T., Maninda, S., & Smith, T. J. (2022). Understanding the differential impact of children's TV on executive functions: a narrative-processing analysis. *Elsevier*.

Hayes, A. (2024). *Wyzowl*. Retrieved from YouTube Stats: Everything You Need to Know In 2024!: https://www.wyzowl.com/youtube-stats/#:~:text=How%20many%20videos%20are%20uploaded%20to%20YouTube%20every%20day%3F,average%20length%20of%204.4%20minutes

Henderson, J. M., Chanceaux, M., & Smith, T. J. (2009). *The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements*. Journal of Vision.

Itti, L. (2007). CHAPTER 94 - Models of Bottom-up Attention and Saliency. In *Neurobiology of Attention* (pp. 576 - 582). Academic Press.

Kirbas, S. I., & Smith, T. J. (2018). How Infants Perceive Animated Films. *BIROn - Birkbeck Institutional Research Online*.

Meta AI. (2024, July). *Meta*. Retrieved from https://ai.meta.com/research/publications/sam-2-segment-anything-in-images-and-videos/

Mital, K. P., Smith, T. J., Hill, R. L., & Henderson, J. M. (2010). *Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion*. Springer.

Murphy, H. (2024). *Financial Times*. Retrieved from https://www.ft.com/content/c581fb74-8d85-4c08-8a46-a7c9ef174454

Parag K. Mital, T. J. (2010). *Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion*. Springer.

Rai1, A. (2019). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*.

Templeton, A., & Conerly, T. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Anthropic*.

Valkenburg, P. M., & Cantor, J. (2002). The Development of a Child into a Consumer. *Journal of Applied Developmental Psychology*.