Appendix

Base Prompt: Zero-shot prompt with CoT Reasoning

A video can be considered inappropriate for children if it contains physical violence (cartoonish, realistic, or gory), interpersonal violence (bullying, pranks, meanness, belittling, controlling behavior, talking down to others, or manipulation), self-harm or suicide (depictions of harm inflicted on oneself or suicidal thoughts/tendencies), extreme stunts (life-endangering, high-risk activities/challenges that require adult supervision), dangerous products or services (like paintball, airsoft, fireworks, weapons, hunting equipment, graffiti products), scary content (horror, suspense, loud violence, zombies, skeletons, masks, scary clowns, blood, dangerous fire, car crashes, medical procedures or other scary visuals), sexual content (innuendos, sexual behavior, nudity, suggestiveness), intimacy (romantic conduct, couples expressing love, sensual contact, dating, relationships or any expression of love or loss for a significant other, lyrics about or mentions of love or romance, heartbreak), inappropriate clothing (significant skin exposure or sheer clothing outside of appropriate contexts such as sunbathing, swimwear,etc), substance use (mention, depictions, or use of alcohol, drugs, tobacco), gambling (betting, gambling, casinos, online or real world casino games, online trading), offensive language (from mild to severe with words such as "stupid" or "idiot"), stereotypes (racial, ethnic, gender, body shaming), crude humor (inappropriate humor, innuendos), weapons (guns, swords, knives, etc but toy guns such as nerf guns are fine), unsuitable media (ads for violent or adult-themed video games, casino games, games with romance or battle or first person-shooter games), fight sports (boxing, wrestling, martial arts), financial content (difference, rolling spot forex, financial spread betting and synonymous products but banking ads are fine), health and beauty services (cosmetic procedures, weight loss, tanning, piercings, tattoos, cosmetics and makeup (no matter their essential benefits) and other products related to external personal care focused on body image but ads on fashion products and essential hygiene products like sunscreen, shampoo, toothpaste, etc. are fine, however, products like sanitary care and hair removal is inappropriate), ads or sponsored content for consumable food or drink products (regardless of nutritional content), where the product is consumed by a person or shown with branding, packaging, or promotional intent are inappropriate; general content like cooking shows, recipes, or scenes where food is simply present or being eaten without promotion are acceptable, pet/animal food products and healthcare, healthcare or medical topics (reproductive health, mental health, substance abuse, eating disorders, pharmaceuticals, health insurance, any nutritional or dietary supplements), sensitive topics (politics, religion, controversial social issues, astrology, occult, paranormal, religious sermons), privacy or safety risks (services or competitions requiring personal information, social networks for strangers, spy services, or video game skin trading), or any explicit depictions or discussions of death or dying. Consider that you are an expert content moderator. Based on both the visual content and audio, is the video advertisement provided inappropriate for children on this definition? If it is, label it as "inappropriate." If it is not, label it as "appropriate." Justify your conclusion. Also, mention what language is being spoken if there is any.

Make the outputs in JSON format:

```
{
  "label": "appropriate" or
  "inappropriate",
  "justification": "brief explanation
  of the classification",
  "language": "language detected in
  the video"
}
```

Justification for Maximum Grid Count Selection for DAVSP

We analyze model performance across varying maximum image grid limits for ads longer than 120 seconds. As shown in Figure 6, increasing the number of grids leads to consistent but diminishing improvements in accuracy and F1 score. For instance, the jump from 1 to 5 grids results in a clear performance gain, while further increases to 10 and 15 grids yield marginal returns.

Given the inference cost per image (e.g., $0.01 for GPT-4, $0.02 for Claude-4), a maximum of 5 image grids offers a strong trade-off between cost and performance. This value is therefore used as the default in subsequent experiments.

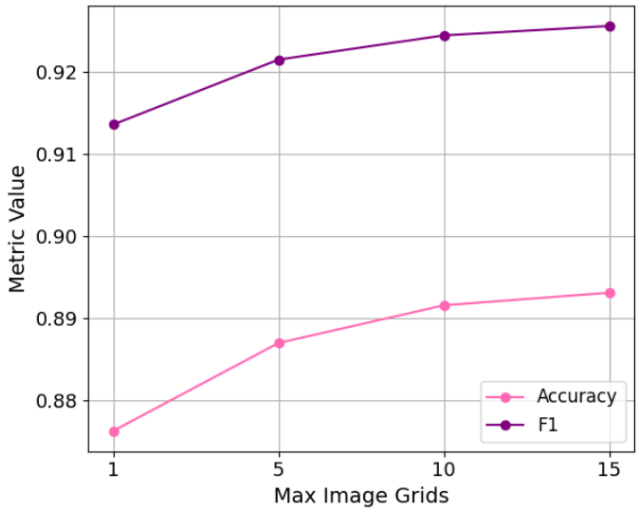

Figure 1: DAVSP: Maximum Number of Image Grids Analysis

Reasons for Invalid Responses

Out of the total 2,423 videos considered for analysis, Gemini failed to generate valid responses for approximately 189 videos. This was due to two non-configurable safety filters: 'Prohibited Content', defined by Gemini as prompts containing restricted material (typically related to CSAM), and 'Other', which encompasses all remaining reasons for prompt rejection. Consequently, these IDs were excluded from the final analysis. Moreover, for cases with metadata where thumbnails were not available, we moved forward without the thumbnail.

Figures and Tables

Language	Count	Language	Count
English	1021	Punjabi	104
French	180	Spanish	93
Swedish	127	Bengali	91
Hindi	125	Arabic	81
Others	125	Urdu	55
German	45	Korean	37
Sinhala	32	Moroccan Arabic	31
Tamil	19	Vietnamese	14
Japanese	14	Turkish	13
Telugu	13	Italian	10
Albanian	10	Haryanvi	10
Malayalam	10	Bhojpuri	9
Portuguese	8	Kannada	5
Swahili	5		

Table 1: Language distribution across the dataset *Note: "Others" includes all languages with fewer than 5 occurrences and videos with no language/linguistic content.*

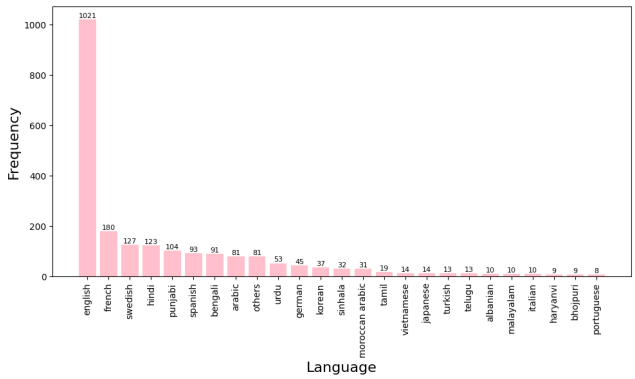

Figure 2: Language Distribution of the dataset (Languages with less than 5 videos were added to others)

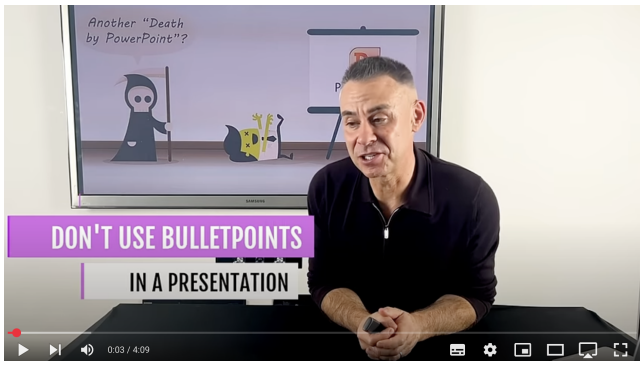

Figure 3: Video & Audio Input: Depiction of Death False Positive (ewQ_wSgCFTU)

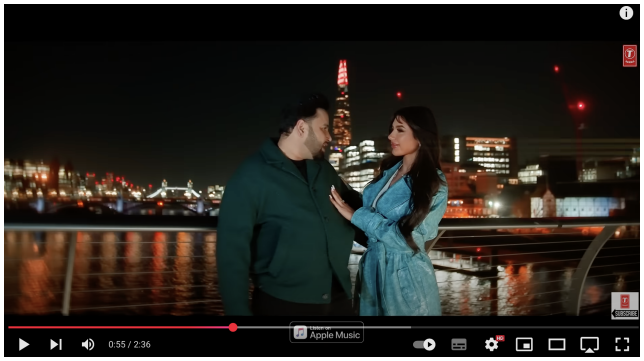

Figure 4: Video & Audio Input: Mild Intimacy False Negative (B1z2h_tZZq0)

Figure 5: Metadata: Fight Sports False Negative (rMnxnd9HvsQ)

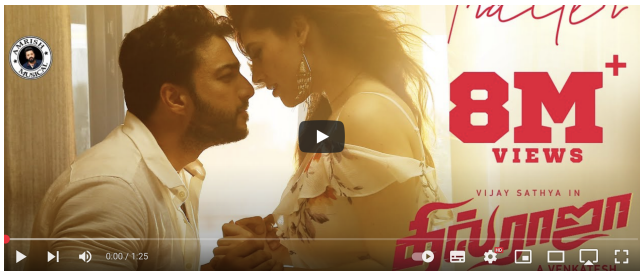

Figure 6: Metadata & English Transcriptions Correctly Classified: Movie Trailer with intimacy and violence initially misclassified by only English Transcriptions input corrected by thumbnail (dN1ZqEuvVGQ)

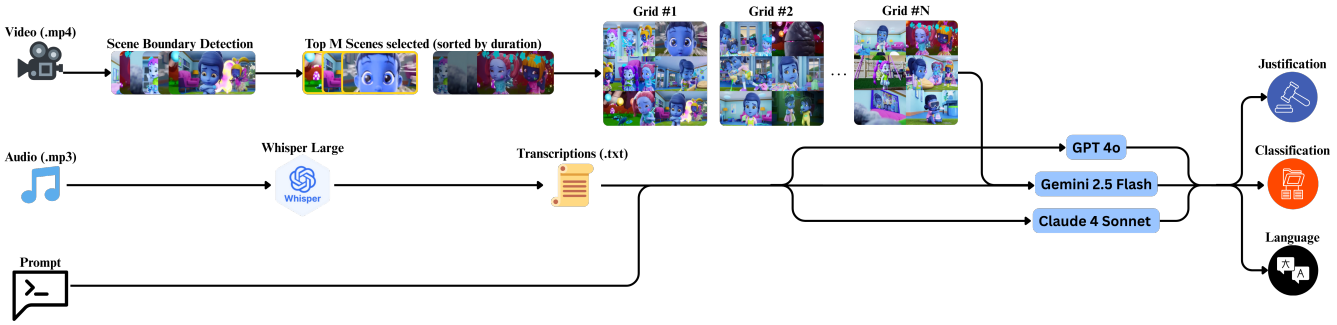

Figure 7: Pipeline Diagram for DAVSP with English Transcriptions

Figure 8: Metadata & English Transcriptions False Negative: Rubber gloves ad with shot of person drinking beer missed due to lack of visual cues (QpnHaKNCaT0)

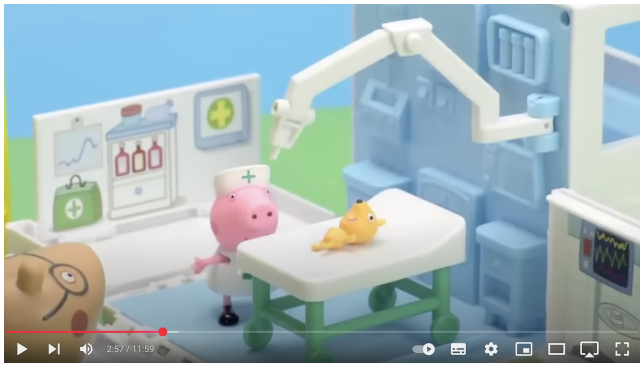

Figure 10: DAVSP & Metadata (Gemini): Peppa Pig X-Ray false positive in DAVSP correctly classified by DAVSP & Metadata (YkVSw2RFGmY)

Figure 9: DAVSP: Gemini Coffee Machine False Positive (6_u0Gsd-guo)

Figure 11: DAVSP & Metadata (Gemini): Domestic Abuse song false negative in DAVSP correctly classified by DAVSP & Metadata (hkworAjntAI)

Figure 12: DAVSP: Claude Food Product False Positive (lLQs7sfJXTw)

Figure 13: DAVSP: GPT Toothpaste False Positive (DwNatxJHBGo)

Figure 14: Example 1: Uniform Sampling

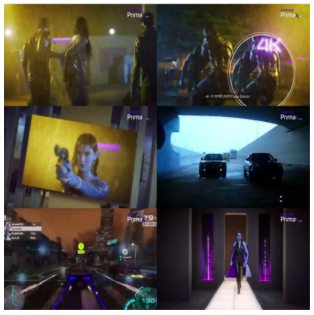

Figure 15: Example 1: Scene Sampling

Figure 16: Example 2: Uniform Sampling

Figure 17: Example 2: Scene Sampling (Image 1)

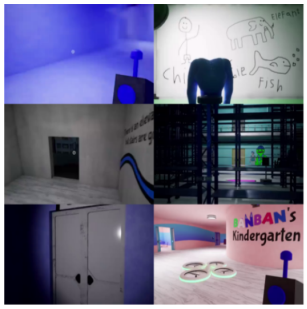

Figure 18: (e) Scene Sampling - Example 2 (Image 2)

Figure 19: Comparison of Uniform Sampling vs Scene Sampling Approaches.

Dataset and Experiment Code Notebooks

The dataset (CSV format), comprising updated labels, secondary tags, transcriptions, and metadata, is included in the Appendix. Each video can be accessed using its ID by appending it to the following URL: https://www.youtube.com/watch?v=\{VideoID\}. Additionally, all experimental notebooks (.ipynb format) used in this study are released alongside the Appendix.

Updated Codebook

Category	Description
Interpersonal violence	Bullying, pranks, meanness, humor at expense of others, belittling ,controlling, talking down to, or manipulating others
Self-harm and Suicide	Depictions of harm inflicted on oneself or suicidal thoughts/tendencies
Scariness	Horror, scary themes, creepy or suspenseful themes, loud and violent content. zombies, skeletons, masks, scary clowns, blood. Fire, mentally distressing scenes (car crashes, medical procedures etc).
Extreme stunts and dangerous content	Depictions of life endangering or high-risk physical activities/challenges Ads for products or services that are dangerous and inappropriate for children or that generally require adult supervision, such as paintball, airsoft, ax-throwing, spray paint, glass-etching substances or graffiti products, fireworks, weapons or weapon accessories, offers related to hunting, etc.
Sexual content	Innuendos, overt sexual behavior, nudity, sexual themes, sexual suggestiveness
Stereotypes	Racial/ethnic, gender, ability, body shaming
Offensive language	Mild (“idiot,” “stupid”) to severe swearing, hate speech or signs
Drinking, drugs, smoking	Mentions, depictions, or use of alcohol, drugs, or tobacco
Unsuitable media	Ads for video games are prohibited if the industry rating of the game would not be suitable for audiences 12 years or under, or if they fall under any of the categories below: Casino Games: Games strongly associated with casinos and betting (even if there isn’t actual money at stake) or guides on how to succeed at such games Adult-themed games: Games featuring sexual elements or other adult content Promotions for games which feature characters kissing or otherwise engaged in romantic pursuit First-person Shooter and Battle games: Games where user has a first-person perspective and shoots at other characters Games that involve controlling armies or fighters in battles or in combat Offers for video game weaponry or character skins Ads for films, TV shows and console games that are sensitive to show to children.
Gambling	Depictions of betting, gambling, casinos, online or real world casino games
Crude humor	Inappropriate humor, innuendos
Display of weapons	Guns, swords, knives
Inappropriate clothing/Significant Skin Exposure	Ads featuring men or women with significant exposed skin or sheer clothing (whether or not the imagery has sexual connotations) are prohibited. Not including running, swimwear, sunbathing etc
Death	

Table 2: Updated Codebook with category definitions as per YouTube’s advertising policies (Part 1)

Category	Description
Intimacy	Sensual physical contact, dating and relationships, apparent romantic conduct between subjects/actors
Fight Sports	Offers related to boxing, wrestling, martial arts and self-defense training.
Financial Content	Contracts for difference, rolling spot forex, financial spread betting and synonymous products.
Health and Beauty	Cosmetics and other products related to external personal care focused on body image. Ads for body modification products or services such as cosmetic procedures, weight loss, tanning, piercings and tattoos. Ads for products related to consumable food and drinks, regardless of nutritional content. Offers related to healthcare and medical issues of all kinds, including reproductive health, substance abuse or recovery, eating disorders, 'miracle cures' and health insurance. Ads for pharmaceuticals or medications, vitamins and nutritional supplements.
Sensitive and Controversial Content	If videos included religious sermons recorded in a place of worship, we did not code for positive or negative content because we did not want to misinterpret religious symbolism. This includes overtly religious content that may contain sermons, worship and other religious rituals Ads for content relating to astrology, the occult or the paranormal. Ads related to politics, religion or other sensitive or controversial social issues.
Privacy/safety/gimmicks	Competitions or sweepstakes promotions, even if free to enter. Offers that require the user to enter their mobile phone number to access content or subscribe to a service. Offers for social networks that allow users to connect with friends and others online. Offers for services that imply they will help spy on a partner, or find non-shared personal information about a third party. Also included are services that perform public records searches for arrest records. Personality quizzes that require entering personal information like an email address or phone number to access results. Ads that sell or promote the trading of video game skins or loot boxes. Offers for platforms or services intended for adults that primarily exist to allow users to connect and communicate with strangers.

Table 3: Updated Codebook with category definitions as per YouTube's advertising policies (Part 2)

Table 4: **Gemini 2.5 Flash Cost Comparison:** Full Video Input vs. DAVSP + Metadata + Transcriptions

Pipeline	Prompt	Image	Transcr.	Audio	Metadata	Video	Total Tokens	Cost ($)
DAVSP + Metadata + Transcr.	720	516	431	0	563	0	2,230	0.000669
Full Video + Audio	720	0	0	4,064	0	33,401	38,185	0.014300