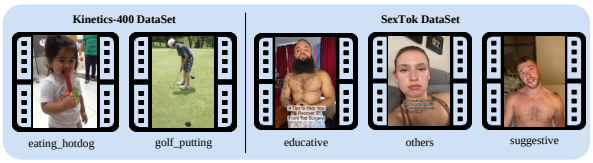
Consensus-aware Balance Learning for Sexually Suggestive Video Classification

Di Zhou[1,2], Jiahui Li[1], Haiying Wang[2], Matthew Burns[2], Meng Liu[1]
[1] Shan Dong Jian Zhu University
[2] Ulster University
{zhoud12222, Ljhh020104, mengliu.sd}@gmail.com
{HY.WANG, M.BURNS2}@ulster.ac.uk

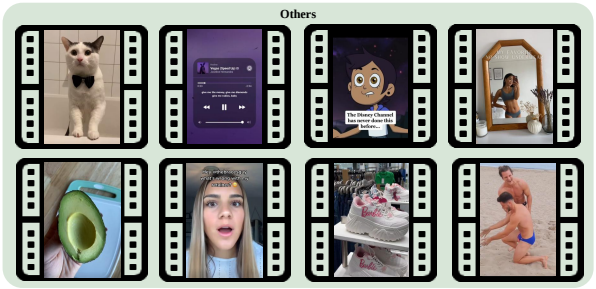
Abstract

In recent years, discussions surrounding sex education have gained considerable attention, as the lack of comprehensive sex education has been linked to various societal issues. While micro-video platforms offer new opportunities for disseminating sex education content, they have also contributed to the proliferation of sexually suggestive videos. Existing video classification methods face significant challenges in this context, such as the difficulty of abstract concepts, cross-domain variation, and training bias due to class imbalance. To address these challenges, we propose a method for classifying sexually suggestive videos. Our approach introduces a consensus-aware visual encoderto assist the model in focusing on the common features of videos within the same category at both the distribution and feature levels, while effectively filtering out irrelevant visual distractions. This improves the model's ability to capture abstract and complex features.Additionally, we employ a label distribution-aware training strategy that allocates more learning capacity to tail classes, ensuring balanced learning across all categories. Experimental results on the SexTok dataset demonstrate that our method excels in classifying sexually suggestive videos, offering improved handling of abstract and imbalanced video content.

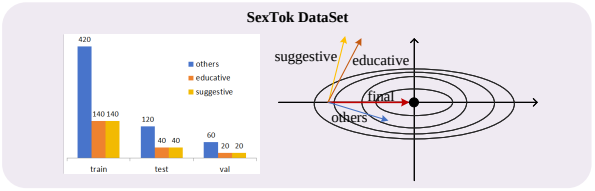
Keywords: Sexually suggestive videos, Video classification, Imbalanced learning, Consensus-aware learning



(a) Abstract Concept



(b) Cross-domain Variation



(c) Class Imbalance

Figure 1. Examples from the SexTok dataset illustrate the challenges of sexually suggestive video classification from three key aspects: abstract concepts, cross-domain variation, and class imbalance.

1. Introduction

In recent years, the discourse surrounding sexual education has gained significant attention as a pressing societal issue. The lack of comprehensive sex education has been linked to both psychological and physical health concerns, alongside the emergence of various social challenges. Addressing these gaps through the promotion of sexual knowledge and reducing the stigma around sexuality is critical for improving individual well-being and fostering societal progress [4].

With the advent of the digital era, micro-video platforms such as Snapchat, TikTok, and Instagram have emerged as pivotal ways of disseminating sexual education. These platforms, known for their concise content, diverse formats, and global reach, offer unique opportunities to deliver educational materials across geographical, cultural, and contextual boundaries, capturing the attention of diverse audiences worldwide [13]. However, the accessibility and open na-

ture of these platforms, coupled with minimal oversight in content creation, have also led to the proliferation of sexually suggestive material. The lack of stringent regulations around video context, production, and presentation presents a significant challenge in controlling the spread of sexually suggestive content. However, in current video context understanding tasks, such as action recognition [29, 38, 5, 39], video classification[27, 34, 26], video anomaly detection [40, 41, 42], and sentiment analysis [44, 21], there is limited focus on distinguishing between suggestive and sex education content. As a result, researchers are increasingly focusing on developing robust strategies to mitigate the spread of such material while maintaining the educational potential of these platforms.

Existing video classification methods typically focus on extracting intra-modal differences to distinguish between video categories. However, these methods face substantial challenges when applied to the classification of sexually suggestive content due to several key factors:

1. **Abstract Concept**: Traditional video datasets [2] [25] [10] [6] [19], such as Kinetics-400 [3] in Figure 1 (a), allow for straightforward classification based on clear, observable actions and scenes. In contrast, the SexTok dataset [17] presents videos where the depicted objects and scenes often lack direct semantic connections to the labels. As illustrated in Figure 1(a), videos with distinct labels may feature creators in similar revealing clothing and similar indoor settings, devoid of distinguishing background features. This highlights a fundamental difficulty: sexually suggestive content, sex education, and general videos often rely on abstract concepts without concrete actions or clear visual distinctions, leading to significant overlap in their visual characteristics.

2. **Cross-domain Variation**: The "others" category in the SexTok dataset encompasses a wide range of content, including daily activities, virtual effects, and animal behavior. As illustrated in Figure 1(b), this category exhibits significant visual diversity, which makes it challenging to extract consistent and discriminative features. The broad scope of content within this category complicates the classification process by increasing the variability of features that the model must handle, thereby making accurate categorization more difficult.

3. **Class Imbalance**: The number of general videos significantly outweighs that of sexually suggestive and sex education videos, leading to class imbalance during training, as illustrated in Figure 1(c). This imbalance causes the model to overfit on the common categories, while underrepresenting rare categories, which hinders the model's ability to capture distinguishing features of minority classes. As a result, classification accuracy and generalization performance deteriorate, posing a persistent challenge for effective video classification in highly imbalanced datasets.

To address these challenges, we propose a novel method for sexually suggestive video classification, built on two key components: 1) **Consensus-aware Visual Encoder**: To overcome the challenges posed by abstract visual expressions in suggestive and educative videos, as well as the wide cross-domain diversity of "others" category videos, we design a consensus-aware visual encoder. It leverages the audio modality as an auxiliary source of information to complement the visual modality. It introduces two sub-modules: multimodal distribution consistency learning and multimodal feature consistency learning. These submodules work at both the distribution and feature levels to minimize interference from irrelevant abstract features in the visual modality, thereby improving the extraction of consistent features within category and discriminative features across categories. 2) **Label Distribution-aware Training Strategy**: To tackle class imbalance, we devise a training strategy that dynamically adjusts the learning process based on label distribution. This strategy allocates additional learning capacity to tail classes (i.e., categories with fewer examples) by modifying class boundaries, allowing the model to learn distinguishing features from underrepresented categories. By optimizing learning across all classes, this approach reduces overfitting to dominant categories and ensures more balanced model performance across the dataset. We conduct experiments on datasets containing sexually suggestive videos, and the experimental results demonstrate that our method significantly outperforms existing approaches in classifying such content. The results highlight the exceptional effectiveness of our approach in overcoming challenges related to abstract concepts, cross-domain variation, and class imbalance, thereby showcasing its robustness and improved performance in this complex classification task.

The key contributions of this paper are as follows:

- We introduce a consensus-aware visual encoder, which enhances the representation of the visual modality to effectively tackle the unique challenges posed by sexually suggestive video classification.

- We propose a label distribution-aware training strategy that addresses class imbalance by allocating additional learning capacity to tail classes, promoting balanced learning across all categories during training.

- Our method outperforms baseline approaches on the SexToK dataset, demonstrating its robustness and effectiveness in classifying sexually suggestive videos.

2. Related Work

With the rapid expansion of internet technology and the increasing openness of social platforms, preventing the dissemination of pornographic and obscene content in online environments has become a critical research focus.

Early work primarily centered on explicit content detection, with an emphasis on nudity detection approaches. Many models [12, 37, 30, 15, 23, 14] rely on segmenting skin-colored regions to identify nudity. However, while these methods can detect overt explicit behaviors, they struggle with more nuanced forms of suggestive content. Another prevalent approach is the bag of visual words model [8, 28, 36, 43], which addresses the semantic gap between low-level visual features and high-level concepts related to explicit content. These models use a collection of image features (visual "words") to capture the visual structure of explicit imagery and achieve more accurate detection. Motion-based analysis techniques have also been explored, capturing movement features in videos to determine inappropriate content. For example, Rea et al. [32] used motion periodicity to identify inappropriate content, while Zuo et al. [45] developed a multimodal detection framework using a Gaussian mixture model to analyze pornographic sounds, combined with contour-based image recognition for visual detection. The final decision is made by integrating both visual and audio inputs. Despite the progress made by these methods, most focus on overtly explicit behaviors characterized by significant skin exposure or large, noticeable movements.

Current detection approaches tend to overlook more subtle, suggestive behaviors, where individuals may be fully clothed or exhibit minimal movement. Moreover, reliance on skin exposure as a criterion for sexual suggestiveness can lead to high false positive rates, especially in contexts such as beachwear or bikinis. The challenge becomes even more complex when suggestive behaviors coexist with educational content, such as sexual health education videos. Existing methods struggle to differentiate between suggestive and educative material, underscoring the need for more sophisticated techniques to detect subtle suggestive behaviors without conflating them with educational content.

3. Methodology

As illustrated in Figure 2, our model architecture is composed of two primary components: consensus-aware visual encoder and label distribution-aware training. The following sections will provide a detailed explanation of the design and functionality of each component.

3.1. Consensus-aware Visual Encoder

To mitigate the interference caused by abstract visual representations in suggestive and educative videos, as well as the diverse visual expressions resulting from the broad cross-domain nature of "others" videos, we introduce the naturally occurring audio modality as additional supervisory information. To leverage this, we design a consensus-aware visual encoder that incorporates two key components: multimodal distribution consistency learning and multimodal feature consistency learning. These components operate at both the distribution and feature levels, enhancing the model's ability to extract informative and robust visual representations by aligning audio and visual features. This approach helps the model focus on meaningful patterns while reducing the impact of irrelevant visual cues.

3.1.1 Multimodal Feature Extraction

Let the untrimmed video be denoted as $\mathcal{V} = \{f_1, \ldots, f_t, \ldots, f_T\}$, where f_t represents the t-th frame and T denotes the total number of frames. We utilize a visual feature extractor (CLIP) [31], pre-trained on a large-scale image dataset (ImageNet) [7], to derive the visual embeddings $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_t, \ldots, \mathbf{v}_T]$. These embeddings are projected into a D-dimensional space using a fully connected (FC) layer with a ReLU activation function. To enhance the model's capability to capture temporal dynamics, we employ a Long Short-Term Memory (LSTM) [20] to obtain both the global visual feature $\mathbf{v}^g \in \mathbb{R}^D$ and local visual features $\bar{\mathbf{V}} = [\bar{\mathbf{v}}_1, \ldots, \bar{\mathbf{v}}_t, \ldots, \bar{\mathbf{v}}_T] \in \mathbb{R}^{T \times D}$. Finally, we fuse the global and local visual features to generate the combined visual feature $\widehat{\mathbf{v}}$:

$$\widehat{\mathbf{v}} = \mathbf{v}^g + \mathcal{F}_{avg\_pooling}(\bar{\mathbf{V}}), \tag{1}$$

where $\mathcal{F}_{avg\_pooling}(\cdot)$ represents the average pooling function, which aggregates the local features across all frames to contribute to the final representation.

The separated audio signal from the untrimmed video $\mathcal{V}$, and then we utilize the audio feature processor AST [18], pre-trained on AudioSet [16], it would be to extract the audio embeddings $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_t, \ldots, \mathbf{a}_T]$, where $\mathbf{a}_t \in \mathbb{R}^D$ and T represents the sequence length of the audio embeddings. These embeddings are then projected into D-dimensional space using a FC layer with a ReLU activation function. To ensure consistency with the visual feature extracting, we apply another LSTM to model the temporal dynamics of the audio, extracting both the global audio feature $\mathbf{a}^g \in \mathbb{R}^D$ and local audio features $\bar{\mathbf{A}} = [\bar{\mathbf{a}}_1, ..., \bar{\mathbf{a}}_t, ..., \bar{\mathbf{a}}_T^s] \in \mathbb{R}^{T \times D}$. Ultimately, the global and local audio features are fused to form the combined audio feature representation $\widehat{\mathbf{a}}$, providing robust support for subsequent multimodal information fusion:

$$\widehat{\mathbf{a}} = \mathbf{a}^g + \mathcal{F}_{avg\_pooling}(\bar{\mathbf{A}}), \tag{2}$$

where $\mathcal{F}_{avg\_pooling}(\cdot)$ represents the average pooling function, which aggregates the local features across all frames to contribute to the final representation.

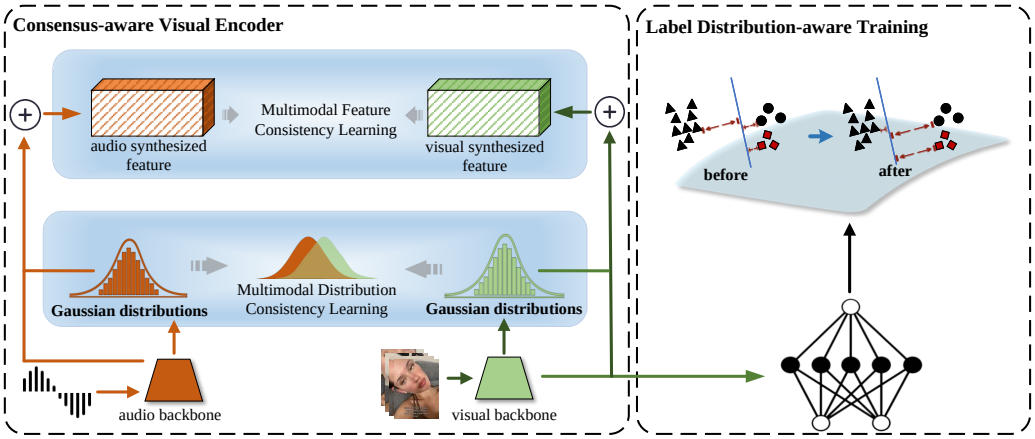


Figure 2. Overview of the proposed Commonalities-Aware Learning Video Classification Network.

3.1.2 Multimodal Distribution Consistency Learning

The goal of the multimodal distribution consistency learning module is to reduce the macro-level differences between the visual and audio modalities. Specifically, both modalities are modeled as Gaussian distributions. To represent these distributions, we define two statistical functions, $\mathcal{G}_v(\cdot)$ and $\mathcal{G}_a(\cdot)$. These functions are implemented using one FC layer followed by ReLU activation functions to compute the parameters of the Gaussian distributions for the visual and audio modalities, denoted as $\mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\sigma}_v^2)$ and $\mathcal{N}(\boldsymbol{\mu}_a$, respectively. The parameters $\boldsymbol{\mu}_*$ and $\boldsymbol{\sigma}_*^2 \in \mathbb{R}^K$ represent the mean and variance of each modality, where K denotes the dimensionality of the latent space. The relationships are formulated as follows:

$$\begin{cases} \mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\sigma}_v^2) = \mathcal{G}_v(\widehat{\mathbf{v}}), \\ \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\sigma}_a^2) = \mathcal{G}_a(\widehat{\mathbf{a}}). \end{cases} \tag{3}$$

To align the Gaussian distributions of the visual and audio modalities, we minimize the divergence between them by optimizing the Kullback-Leibler (KL) divergence. Specifically, we aim to reduce the KL divergence between $\mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\sigma}_v^2)$ and $\mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\sigma}_a^2)$, effectively minimizing the differences between the distributions. The objective function for this optimization is defined as follows:

$$\mathcal{L}_{KL} = -\frac{1}{2}\sum_{k=1}^{K}[log(g_k) - g_k - \gamma_k + 1], \tag{4}$$

where $\mathbf{g} = \frac{\boldsymbol{\sigma}_v^2}{\boldsymbol{\sigma}_a^2}$ and $\boldsymbol{\gamma} = \frac{(\boldsymbol{\mu}_v - \boldsymbol{\mu}_a)^2}{\boldsymbol{\sigma}_v^2}$, K is the dimension of the latent space, and $*_k$ denotes the k-th element of the corresponding vector. This optimization ensures that the visual and audio feature distributions are statistically aligned, enabling better multimodal consistency.

3.1.3 Multimodal Feature Consistency Learning

The goal of multimodal feature consistency learning is to further align the visual and audio modalities at the feature level by reducing their discrepancies. This module ensures the alignment of internal similarities between the visual and audio feature distributions, facilitating a precise alignment in feature representation.

We begin by employing a reparameterization technique to generate synthesized features $\mathbf{v}^r \in \mathbb{R}^D$ for the visual modality and $\mathbf{a}^r \in \mathbb{R}^D$ for the audio modality:

$$\begin{cases} \mathbf{v}^r = \mathcal{F}_{map_v}(\boldsymbol{\mu}_v + \boldsymbol{\sigma}_v^2 \delta_I), \\ \mathbf{a}^r = \mathcal{F}_{map_a}(\boldsymbol{\mu}_a + \boldsymbol{\sigma}_a^2 \delta_I), \end{cases} \tag{5}$$

where $\delta_I \sim \mathcal{N}(0, \mathbf{1})$ follows a normal distribution, and $\mathcal{F}_{map_a}(\cdot)$ and $\mathcal{F}_{map_v}(\cdot)$ are mapping functions that project the synthesized features from $\mathbb{R}^K$ to $\mathbb{R}^D$. These mapping functions consist of an FC layer followed by a ReLU activation function for both visual and audio modalities.

Next, we compute the intra-modal feature similarity matrices for the visual and audio modalities by calculating the cosine similarity between each pair of videos. For the visual modality, the similarity score $s_{i,j}^v$ between the i-th and j-th videos is given by:

$$s_{i,j}^v = \left(\frac{\mathbf{v}_i^r + \widehat{\mathbf{v}}_i}{||\mathbf{v}_i^r + \widehat{\mathbf{v}}_i||_2}\right)^T \left(\frac{\mathbf{v}_j^r + \widehat{\mathbf{v}}_j}{||\mathbf{v}_j^r + \widehat{\mathbf{v}}_j||_2}\right), \tag{6}$$

and for the audio modality, the similarity score $s_{i,j}^a$ is:

$$s_{i,j}^a = \left(\frac{\mathbf{a}_i^r + \widehat{\mathbf{a}}_i}{||\mathbf{a}_i^r + \widehat{\mathbf{a}}_i||_2}\right)^T \left(\frac{\mathbf{a}_j^r + \widehat{\mathbf{a}}_j}{||\mathbf{a}_j^r + \widehat{\mathbf{a}}_j||_2}\right). \tag{7}$$

Finally, to achieve precise feature-level alignment between the visual and audio modalities, we minimize the distance between the visual modality similarity matrix $\mathbf{s}^v =$

$[s^v_{1,1}, \ldots, s^v_{1,j}, \ldots, s^v_{i,j}, \ldots, s^v_{B,B}]$ and the audio modality similarity matrix $\mathbf{s}^a = [s^a_{1,1}, \ldots, s^a_{1,j}, \ldots, s^a_{i,j}, \ldots, s^a_{B,B}]$. This optimization ensures that the feature representations of the two modalities are consistent:

$$\mathcal{L}_{sim} = \frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} (s^v_{i,j} - s^a_{i,j})^2, \tag{8}$$

where B represents the size of the minibatch. This approach ensures that visual and audio features are closely aligned at the feature representation level, improving multimodal consistency and enhancing the model's ability to extract meaningful cross-modal information.

3.2. Label Distribution-aware Training

After passing through multimodal distribution consistency learning and multimodal feature consistency learning, the visual feature $\widehat{\mathbf{v}}$ has absorbed the semantic information from the audio modality and exhibits enhanced discriminative ability. Therefore, in the final prediction stage, we design a simple yet effective classification network to process these visual features. The visual feature $\widehat{\mathbf{v}}$ is passed through the classification network, $\mathcal{F}_{MLP}(\cdot)$, which consists of two FC layers and a ReLU activation function. This network analyzes the visual features comprehensively and computes the class probabilities $\mathbf{p} = [p_1, ..., p_c, ..., p_C]$, where C represents the number of categories in the dataset. The final probability for the c-th class is calculated as follows:

$$\mathbf{p} = \mathcal{F}_{MLP}(\widehat{\mathbf{v}}). \tag{9}$$

To improve the generalization ability for tail classes and mitigate the adverse effects of class imbalance during model training, In other words, by increasing the minimum distance between long-tail class samples and the decision boundary, the classification margin is expanded, effectively reducing the generalization error for long-tail classes. First, the total number of samples for each class in the dataset is computed and represented as $\mathcal{C}_{class} = \{Num_c\}_{c=1}^{C}$, where Num_c represents the total number of samples in the c-th class. We then define a hyperparameter θ as the maximum boundary value, and the boundary value $\overline{\theta}_c$ for each class is calculated as follows:

$$\begin{cases} \theta_c = \dfrac{1}{\sqrt[4]{Num_c}}, \ c \in \{1, \ldots, C\} \\ \overline{\theta}_c = \dfrac{\theta * \theta_c}{max\{\theta_1, \theta_2, ..., \theta_C\}}, \ c \in \{1, \ldots, C\}. \end{cases} \tag{10}$$

Finally, we incorporate the margin coefficient into the standard cross-entropy loss function, resulting in a label distribution-aware balanced cross-entropy loss that is formulated as:

$$\mathcal{L}_{balance} = -log[\frac{e^{\beta \cdot (p_c - \overline{\theta}_c)}}{e^{\beta \cdot (p_c - \overline{\theta}_c)} + \sum_{\widehat{c} \neq c} e^{p_{\widehat{c}}}}], \tag{11}$$

where β is an adjustable hyperparameter, and $p_*, * \in \{1, ..., c, ..., C\}$ represents the class score assigned to the sample by the model, are learned through the classification network. This strategy expands the decision boundary for tail class samples by increasing the minimum distance between the decision boundary and tail class samples, thus providing a larger margin for these classes. This approach enhances classification performance on tail classes while improving overall generalization.

The overall loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{balance} + \lambda_1 \mathcal{L}_{KL} + \lambda_2 \mathcal{L}_{sim}, \tag{12}$$

where λ_1 and λ_2 are hyperparameters to weigh the contributions of different loss functions.

4. Experiment

4.1. Datesets

The SexTok dataset, introduced by George et al. [17], is currently the only publicly available dataset for the task of sexually suggestive video classification. Specifically curated from the TikTok platform, the dataset categorizes videos into three distinct classes: suggestive, educative, and others, with the latter covering a wide range of content. The dataset consists of 1,000 videos with durations ranging from 1 second to 7 minutes. The video content is diverse, spanning everyday life, educational, and entertainment contexts. This dataset provides a valuable resource for exploring this emerging task.

For evaluating the performance of classification models, the SexTok dataset is divided into training, validation, and test sets. Notably, the dataset exhibits a significant class imbalance, as sexually suggestive and sex education content are substantially outnumbered by other videos. Specifically, sexually suggestive videos make up around 20%, sex education videos account for another 20%, and the remaining 60% fall under the others video category. This imbalance poses challenges for model training and optimization, making the dataset a suitable benchmark for testing the effectiveness of multimodal alignment and class balancing strategies.

4.2. Evaluation Metrics

We performed a comprehensive evaluation of our model and baseline methods using a variety of assessment metrics on the SexTok dataset. The evaluation metrics include accuracy, precision, recall, and F1 score to provide a standard comparison of classification performance. We further refined the evaluation by calculating micro-precision, micro-recall, and micro-F1 scores, where we treated the "other" category as the negative class and excluded it from the final score. This approach allows for a more focused evaluation of the model's performance on the inappropriate content classes. Additionally, we computed macro-precision,

Table 1. Performance comparison with several state-of-the-art baselines on the SexTok dataset. The best performance is highlighted in **bold**.

Methods	Accuracy	Micro			Macro		
		Precision	Recall	F1	Precision	Recall	F1
All-text Bert [17]	68%	76%	50%	60%	71%	63%	64%
Non-empty Text Bert [17]	75%	78%	54%	64%	74%	65%	68%
Visual-VideoMAE [17]	70%	61%	51%	55%	68%	57%	61%
Slowfast [11]	80%	95%	63%	76%	81%	73%	76%
Timesformer [1]	75%	93%	52%	66%	75%	65%	68%
ResNet [22]	77%	90%	75%	63%	77%	64%	67%
Uniformer [24]	74%	93%	55%	69%	73%	66%	68%
Ours	**86%**	**97%**	**81%**	**88%**	**85%**	**84%**	**84%**

Table 2. Performance comparison with several state-of-the-art baseline models for overall F1 performance across each category label. The best performance for each category is highlighted in **bold**.

Methods	Suggestive	Educative	Others
All-text Bert	30%	83%	80%
Non-empty Text Bert	38%	84%	81%
Visual-VideoMAE	55%	63%	72%
Slowfast	74%	68%	86%
Timesformer	67%	55%	83%
ResNet	64%	54%	84%
Uniformer	71%	55%	80%
Ours	**81%**	82%	**90%**

macro-recall, and macro-F1 scores, which offer a better understanding of the model's effectiveness across all categories, regardless of class size, thus addressing the challenges posed by the imbalanced data distribution.

4.3. Implementation Details

Our model was optimized using a single GeForce RTX 2080 Ti GPU and implemented with the PyTorch library. Frame-wise visual features were extracted using a CLIP model [31] pre-trained on the ImageNet [7] dataset, while audio features were obtained from an AST model [18] pre-trained on the AudioSet [16] dataset.

Given the variation in video lengths within the SexTok [17] dataset, we standardized the sequence length to 800 frames, which represents the median value across the dataset. For videos containing more than 800 frames, we applied a uniform sampling strategy to select 800 frames. For videos with fewer than 800 frames, we used interpolation to expand the sequence to 800 frames. To avoid excessive redundancy in videos that are significantly shorter than 800 frames, which could lead to increased computational overhead, we implemented a sliding window approach. Specifically, we used a window size of 16 frames with a stride of 8, followed by average pooling within each window. This process standardizes all videos to 100 frames.

For model optimization, we used the AdamW optimizer with a mini-batch size of 16 and a learning rate of $6e-5$. The balance parameters λ_1 and λ_2 were set to 0.01 and 0.9, respectively, while the hyperparameter s was fixed at 30. During the experiments, the model with the best test performance was selected as the final model for evaluation.

4.4. Baselines

To demonstrate the effectiveness of our proposed model, we conducted a comprehensive benchmarking against several state-of-the-art models. This comparison included BERT [17][9], configured for both All-text and Non-empty text scenarios, Visual-VideoMAE [17][35], SlowFast [11], TimeSformer [1], ResNet [22], and UniFormer [24]. The benchmarking was carried out on the SexTok dataset. With the exception of BERT and Visual-VideoMAE, the other baseline methods were constructed by freezing the model parameters, utilizing them as powerful visual feature extractors, and attaching a classifier on top. Below is a detailed explanation of these baseline methods:

- **All-text Bert** and **Non-empty Text Bert**: We followed the model configuration used by George et al. [17] to fine-tune the BERT-base-multilingual-cased model [9] for classifying text transcripts from videos. These transcripts were generated by converting audio information into text using OpenAI's Whisper (medium) [33]. Since some videos consist solely of music or lack audio, resulting in empty transcripts, we implemented two distinct setups: the All-text BERT model, which includes all video transcriptions, and the Non-empty Text BERT model, which excludes videos without text.

- **Visual-VideoMAE**: In line with George et al. [17], we fine-tuned the MCG-NJU/videomae-base model (Tong et al., 2022), which is designed for video classification tasks. The video data was preprocessed and sampled using the same strategies as George et al. to ensure consistency in evaluation.

- **Slowfaster**: SlowFast[11] is a dual-pathway model that processes video by integrating both temporal and

spatial information. It uses a Slow pathway for capturing spatial details and a Fast pathway for temporal dynamics. We used the SlowFast model pre-trained on the Kinetics-400 (K400) dataset as a baseline, to assess its performance on the SexTok dataset.

- **Timesformer**: TimeSformer [1] is a Transformer-based model designed for video content, employing a "Divided Space-Time Attention" mechanism to handle spatio-temporal information efficiently. Pre-trained on the Kinetics-400 dataset, TimeSformer served as a baseline model for evaluating performance on the SexTok dataset.

- **ResNet**: ResNet [22] uses a residual learning framework that helps mitigate the vanishing gradient problem, enabling deep networks to remain trainable. A ResNet model pre-trained on ImageNet was used as a baseline to evaluate its generalization in classifying videos within the SexTok dataset.

- **Uniformer**: UniFormer [24] is a hybrid Transformer architecture that combines the strengths of convolutional and self-attention mechanisms for learning spatiotemporal features from video data. Pre-trained on Kinetics-400, the UniFormer model was used as a baseline to evaluate its effectiveness in handling poor-quality video classification and its performance on the SexTok dataset.

4.5. Performance Comparison

The analysis of Table 1 reveals several key observations. Models based on a single visual modality generally outperform those based on a single text modality. Except for Visual-VideoMAE, most visual modality models outperform text-based models across all evaluation metrics, suggesting that the visual modality provides more comprehensive information for sexually suggestive classification tasks. However, in the task of sexually suggestive video classification, despite their strong performance in other tasks, visual modality models fall short compared to our proposed model. This can be attributed to the complexity of abstract labels, the similarity in visual features, and the wide diversity of cross-domain video content, which challenge models relying solely on visual information. Specifically, our model achieves an accuracy of 86%, which is 6% points higher than the next-best model. In the micro evaluation group, our model's precision is 2% points higher than the second-best, with an increase in recall of 6% points, and a 12%-point improvement in F1-score. In the macro group, our model surpassed others by 4% in precision, 11% in recall, and 8% in the F1-score. These results demonstrate our model's superior ability to understand abstract concepts and extract critical information from videos with similar visual representations and diverse cross-domain categories.

4.6. Ablation Study

In this section, we present the results of the ablation study to evaluate the contribution of each module in our proposed model to the overall performance. Specifically, we examined the impact of the consensus-aware visual encoder, label distribution-aware training strategy, and feature selection on sexually suggestive video classification.

4.6.1 On Consensus-aware Visual Encoder

To assess the effectiveness of the consensus-aware visual encoder in understanding abstract labels, addressing the similarity of visual representations, and managing the broad scope of cross-domain videos, we introduced three variations of the model:

- **w/o AUDIO**: In this variation, we removed the audio modality information from the commonalities-aware learning, retaining only the operations on the visual modality. This allows us to verify the effectiveness of using only visual information in the classification task and understand the contribution of the audio modality to performance.

- **w/o DCL**: We eliminated the multimodal distribution consistency learning to assess the importance of aligning the Gaussian distributions of the visual and audio modalities at a macro level. This variation helps quantify the impact of distribution-level alignment on overall classification accuracy and performance.

- **w/o FCL**: In this model variation, we removed the multimodal feature consistency learning to evaluate the necessity of aligning the visual and audio modalities at the feature level. This analysis highlights the contribution of feature-level consistency to the model's ability to generalize across different modalities and video categories.

From Figure 3, we can see that our model consistently outperforms the other variants across various metrics. Specifically, when compared to the w/o AUDIO model, we observed improvements of 6% in Accuracy, 6% in Micro-Recall, 4% in Micro-F1, 3% in Macro-Precision, 4% in Macro-Recall, 3% in Macro-F1, and further gains of 2% in the classification accuracy for the "Suggestive", "Educative", and "Others" categories. Similarly, when compared to the w/o DCL model, we saw improvements of 2% in Accuracy, 1% in Micro-Recall, 3% in Micro-F1, 3% in Macro-Precision, 3% in Macro-Recall, 2% in Macro-F1, 3% in the "Suggestive" category, 3% in the "Educative" category, and 2% in the "Others" category. The same pattern emerged when comparing our model to the w/o FCL variant, with improvements of 2% in Accuracy, 2% in Micro-Recall, 3% in Micro-F1, 3% in Macro-Precision, 3% in Macro-Recall,

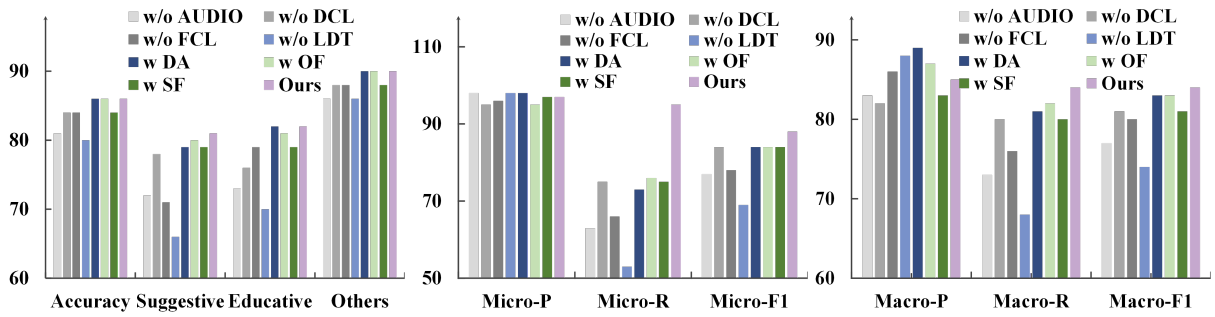


Figure 3. Ablation study results on the SexTok dataset.

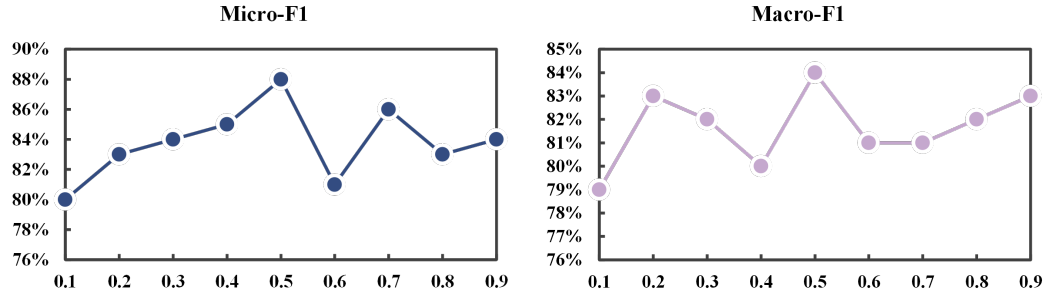


Figure 4. Influence of θ on the classification performance.

2% in Macro-F1, 3% in the "Suggestive" category, 3% in the "Educative" category, and 2% in the "Others" category, respectively.

These results demonstrate that our proposed consensus-aware visual encoder significantly boosts the model's capacity to comprehend abstract concepts, effectively addressing challenges related to visual similarity and the wide diversity of cross-domain video content.

4.6.2 On Label distribution-aware Training

To further emphasize the importance of the label distribution-aware training strategy introduced in our model, we designed two variants of our model:

- **w/o LDT**: In this variant, we removed the label distribution-aware training strategy, thus disregarding the impact of data imbalance on the model. This baseline variant uses standard binary cross-entropy loss without accounting for class imbalance.

- **w DA**: This variant addresses the imbalance by replicating videos from the "suggestive" and "educative" categories to match the data size of the "others" category, thus artificially balancing the dataset through data augmentation.

As illustrated in Figure 3, our model consistently achieves superior performance across multiple metrics. Compared to the w/o LDT model, we observed improvements of 6% in Accuracy, 25% in Suggestive, 17% in Educative, 5% in Others, 4% in Micro-Precision, 9% in Micro-Recall, 4% in Micro-F1, 3% in Macro-Precision, 4% in Macro-Recall, and 3% in Macro-F1. Similarly, when compared to the w DA model, the same metrics exhibited improvements of 2% in Suggestive, 2% in Micro-Recall, 4% in Micro-F1, 4% in Macro-Precision, 3% in Macro-Recall, and 1% in Macro-F1. Notably, the w/o LDT model performs the worst, consistently underperforming across nearly all metrics. This clearly demonstrates that applying a class balancing strategy is critical for enhancing the model's performance in the task of sexually suggestive video classification, particularly in the face of class imbalance. This strategy significantly contributes to improved classification accuracy and generalization across all video categories.

4.6.3 On Feature Selection

To investigate which feature is more beneficial for training the multimodal feature consistency learning module (in Eqn. (6) and Eqn. (7)), we designed two model variants:

- **w OF**: This variant uses only the original visual features for training the model, without incorporating any synthetic features.

- **w SF**: This variant relies solely on synthetic features, generated through Gaussian distribution calculations,

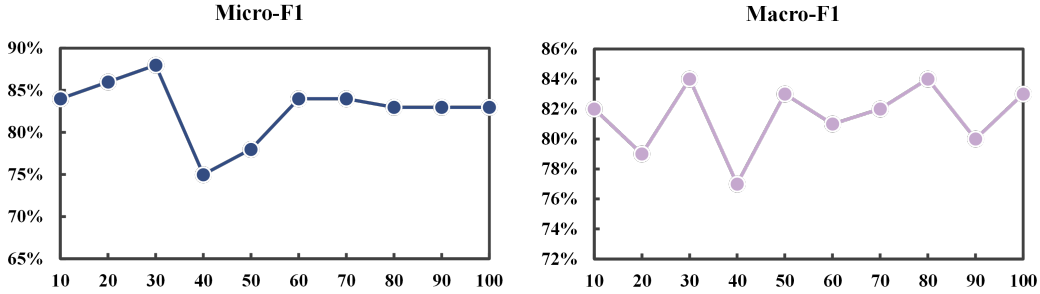


Figure 5. Influence of β on the classification performance.

for training.

As illustrated in Figure 3, our model demonstrates a clear advantage by combining both synthetic and original features instead of using only one type of feature. Compared to the w OF variant, our model shows improvements of 1% in Suggestive, 1% in Educative, 1% in Others, 1% in Micro-Precision, 9% in Micro-Recall, 4% in Micro-F1, 3% in Macro-Precision, 9% in Macro-Recall, and 11% in Macro-F1, respectively. Similarly, when compared to the w SF variant, our model achieves increases of 2% in Accuracy, 2% in Suggestive, 3% in Educative, 2% in Others, 13% in Micro-Recall, 5% in Micro-F1, 2% in Macro-Precision, 6% in Macro-Recall, and 4% in Macro-F1. These results strongly suggest that combining synthetic and original features within the multimodal feature consistency learning module significantly enhances the model's overall performance, allowing it to better capture and align features across modalities. This combined approach proves to be more effective than relying on either type of feature alone.

4.7. Parameter Analysis

The Maximum Boundary θ. To explore the impact of the maximum boundary θ in Eqn. (10), we conducted a parameter analysis experiment on the SexTok dataset under different θ values (ranging from 0.1 to 0.9, with an increment of 0.1), evaluated using Micro-F1 and Macro-F1. The experimental results are shown in Figure 4. The results demonstrate that as the θ value increases, both Micro-F1 and Macro-F1 follow a general trend of rising initially and then declining, reaching their peak at $\theta = 0.5$. Although additional fluctuations occur as θ increases, none of these surpass the optimal performance achieved at $\theta = 0.5$. Therefore, through this parameter analysis, we determined that $\theta = 0.5$ is the optimal value for our proposed model.

The hyperparamete β. To explore the impact of the hyperparameter β in Eqn. (11), we conducted a parameter analysis experiment on the SexTok dataset under different β values (ranging from 10 to 100, with an increment of 10), measured by Micro-F1 and Macro-F1. The experimental results are shown in Figure 5. The results demonstrate that as the β value increases, Micro-F1 follows an overall trend of rising initially, then declining, and finally stabilizing, achieving its best performance at $\beta = 30$. Macro-F1, on the other hand, fluctuates throughout but also reaches its peak at $\beta = 30$. Despite additional fluctuations as $\beta = 30$ increases, none surpass the optimal performance obtained at $\beta = 30$. Therefore, through this parameter analysis, we determined that $\beta = 30$ is the optimal value for our proposed model.

4.8. Qualitative Analysis

In Figure 6, we present representative cases from the SexTok dataset to qualitatively evaluate the effectiveness of our proposed method. Each column corresponds to a specific category, comparing the classification results of our model with the second-best model, SlowFast [11].

In the first row of the first and third columns, the videos exhibit strong explanatory features, with individuals verbally explaining concepts. The second-best model, influenced by these "explanatory" characteristics, incorrectly classified these videos as "educative". In contrast, our model effectively distinguished between surface-level explanatory features and the actual content, resulting in correct classifications.

In the second row of the first and second columns, the videos contain prominent visual cues, such as significant exposure of body parts. The second-best model heavily relied on these obvious visual features, leading to incorrect classifications as "suggestive". However, our model avoided making decisions based solely on visual exposure and correctly identified the underlying content, demonstrating a deeper understanding of the video context.

Furthermore, the first row of the second column and the second row of the third column showcase videos that were accurately classified as "others" by both models. These videos have minimal visual complexity, limited movement, and rely primarily on audio cues for context. Our model was able to correctly capture the audio-dominant nature of these videos, preventing misclassification caused by oversimplified visual cues.

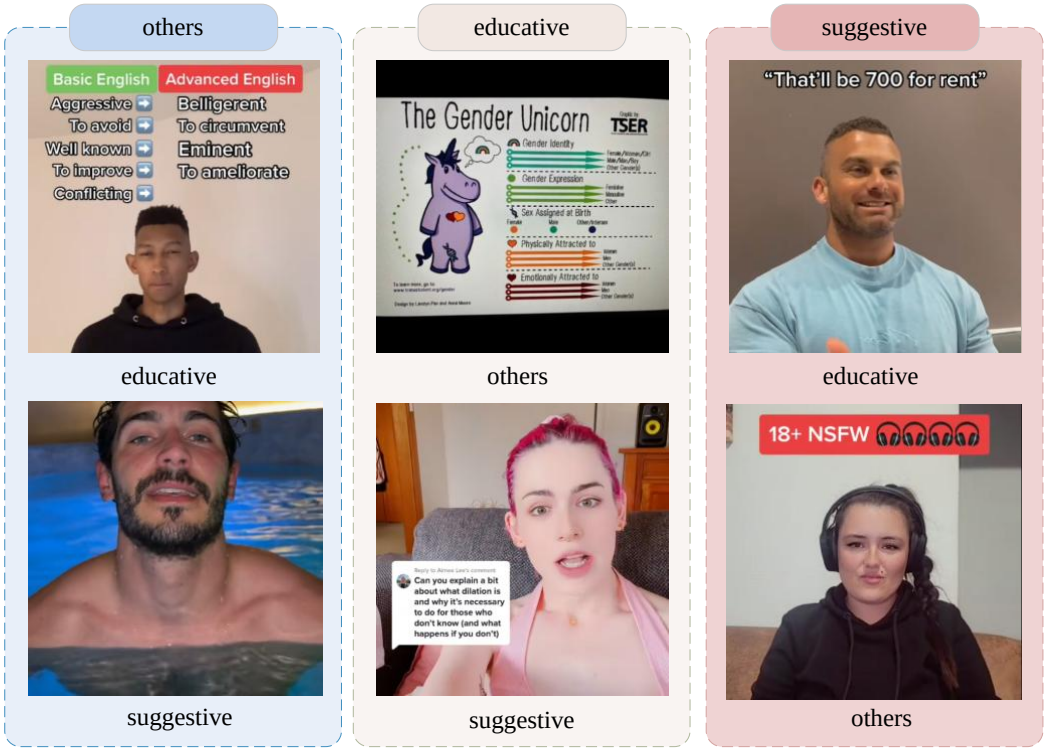


Figure 6. Qualitative comparison of our model with the suboptimal model, SlowFast, on the SexTok dataset. The figure highlights cases where the SlowFast model made incorrect classifications: in the first column, the true label is "others"; in the second column, the true label is "educative"; and in the third column, the true label is "suggestive".

These qualitative analyses highlight the superior performance of our proposed method in handling complex and abstract concepts. Compared to the second-best model, our approach demonstrates greater precision in filtering out irrelevant features, focusing on key information, and enhancing both the accuracy and robustness of sexually suggestive video classification.

5. Conclusion

In this work, we address the challenge of classifying sexual education and sexually suggestive videos by proposing a novel video classification approach. Our method enhances the model's capacity to identify shared features across video content through the introduction of a consensus-aware visual encoder. Additionally, we implement a label distribution-aware training strategy that dynamically adjusts the learning process to provide additional support for underrepresented categories, ensuring balanced learning across all classes. Experimental results on the SexTok dataset demonstrate the effectiveness of our approach.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China, No.:62376140 and No.:U23A20315 ; the Science and Technology Innovation Program for Distinguished Young Scholars of Shandong Province Higher Education Institutions, No.:2023KJ128; the Special Fund for Taishan Scholar Project of Shandong Province; and the Special Fund for Distinguished Professors of Shandong Jianzhu University.

References

[1] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 6, 7

[2] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. A short note about kinetics-600. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Aug 2018. 2

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 2

[4] R. L. Collins, V. C. Strasburger, J. D. Brown, E. Donnerstein, A. Lenhart, and L. M. Ward. Sexual media and childhood well-being and health. *Pediatrics*, 140(Supplement_2):S162–S166, 2017. 1

[5] E. Dastbaravardeh, S. Askarpour, M. Saberi Anari, and K. Rezaee. Channel attention-based approach with autoencoder network for human action recognition in low-

resolution frames. *International Journal of Intelligent Systems*, 2024(1):1052344, 2024. 2

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2009. 2

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 6

[8] T. Deselaers, L. Pimenidis, and H. Ney. Bag-of-visual-words models for adult image classification and filtering. In *2008 19th International Conference on Pattern Recognition*, Dec 2008. 3

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019. 6

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Oct 2020. 2

[11] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 6, 9

[12] M. M. Fleck, D. A. Forsyth, and C. Bregler. *Finding naked people*, page 593–602. Jan 1996. 3

[13] L. R. Fowler, L. Schoen, H. S. Smith, and S. R. Morain. Sex education on tiktok: a content analysis of themes. *Health promotion practice*, 23(5):739–742, 2022. 1

[14] D. Ganguly, M. H. Mofrad, and A. Kovashka. Detecting sexually provocative images. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, volume 3, page 660–668, Mar 2017. 3

[15] M. B. Garcia, T. F. Revano, B. G. M. Habal, J. O. Contreras, and J. B. R. Enriquez. A pornographic image and video filtering application using optimized nudity recognition and detection algorithm. In *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*, volume 521, page 1–5, Nov 2018. 3

[16] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 3, 6

[17] E. George and M. Surdeanu. It's not sexually suggestive; it's educative— separating sex education from suggestive content on tiktok videos. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5904–5915, 2023. 2, 5, 6

[18] Y. Gong, Y.-A. Chung, and J. Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 3, 6

[19] R. Goyal, S. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. Jun 2017. 2

[20] A. Graves and A. Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012. 3

[21] F. Hasan, D. M. Raza, H. Moon, and M. A. H. Nahid. Sentiment analysis from youtube video using bi-lstm-gru classification check for updates. In *Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning: BIM 2023*, volume 867, page 303. Springer Nature, 2024. 2

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[23] H. Lee, S. Lee, and T. Nam. Implementation of high performance objectionable video classification system. In *2006 8th International Conference Advanced Communication Technology*, volume 4, page 4 pp. – 962, Jan 2006. 3

[24] K. Li, Y. Wang, G. Peng, G. Song, Y. Liu, H. Li, and Y. Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *International Conference on Learning Representations*. 6, 7

[25] Y. Li, Y. Li, and N. Vasconcelos. *RESOUND: Towards Action Recognition Without Representation Bias*, page 520–535. Jan 2018. 2

[26] M. Liu, L. Nie, M. Wang, and B. Chen. Towards micro-video understanding by joint sequential-sparse modeling. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 970–978, 2017. 2

[27] M. Liu, L. Nie, X. Wang, Q. Tian, and B. Chen. Online data organizer: Micro-video categorization by structure-guided multimodal dictionary learning. *IEEE Transactions on Image Processing*, 28(3):1235–1247, 2019. 2

[28] A. Lopes, S. Avila, A. Peixoto, R. Oliveira, and A. Araújo. A bag-of-features approach based on hue-sift descriptor for nude detection. *European Signal Processing Conference,European Signal Processing Conference*, Aug 2009. 3

[29] Y. Ma and R. Wang. Relative-position embedding based spatially and temporally decoupled transformer for action recognition. *Pattern Recognition*, 145:109905, 2024. 2

[30] C. Platzer, M. Stuetz, and M. Lindorfer. Skin sheriff. In *Proceedings of the 2nd international workshop on Security and forensics in communication systems*, Jun 2014. 3

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[32] R. Rea, L. Lacey, D. Dahyotit, and D. Dahyot. Multimodal periodicity analysis for illicit content detection in videos. *Conference on Visual Media Production,Conference on Visual Media Production*, Jan 2006. 3

[33] R. W. Summers. *Social Psychology: How Other People Influence Our Thoughts and Actions [2 volumes]*. Bloomsbury Publishing USA, 2016. 6

[34] W. Tan, Q. Yao, and J. Liu. Overlooked video classification in weakly supervised video anomaly detection. 2

[35] Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 6

[36] A. Ulges and A. Stahl. Automatic detection of child pornography using color visual words. In *2011 IEEE International Conference on Multimedia and Expo*, volume 4, page 1–6, Jul 2011. 3

[37] D. Wang, M. Zhu, X. Yuan, and H. Qian. Identification and annotation of erotic film based on content analysis. In *SPIE Proceedings,Electronic Imaging and Multimedia Technology IV*, volume 5637, page 88, Feb 2005. 3

[38] X. Wang, S. Zhang, J. Cen, C. Gao, Y. Zhang, D. Zhao, and N. Sang. Clip-guided prototype modulating for few-shot action recognition. *International Journal of Computer Vision*, 132(6):1899–1912, 2024. 2

[39] X. Wang, S. Zhang, Z. Qing, Z. Zuo, C. Gao, R. Jin, and N. Sang. Hyrsm++: Hybrid relation guided temporal set matching for few-shot action recognition. *Pattern Recognition*, 147:110110, 2024. 2

[40] P. Wu, X. Zhou, G. Pang, Y. Sun, J. Liu, P. Wang, and Y. Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18297–18307, 2024. 2

[41] P. Wu, X. Zhou, G. Pang, L. Zhou, Q. Yan, P. Wang, and Y. Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6074–6082, 2024. 2

[42] Z. Yang and R. J. Radke. Context-aware video anomaly detection in long-term datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4002–4011, 2024. 2

[43] J. Zhang, L. Sui, L. Zhuo, Z. Li, and Y. Yang. An approach of bag-of-words based on visual attention model for pornographic images recognition in compressed domain. *Neurocomputing*, 110:145–152, Jun 2013. 3

[44] J. Zhao, H. Liu, Y. Wang, W. Zhang, X. Zhang, B. Li, T. Sun, Y. Qi, and S. Zhang. Sentiment analysis of video danmakus based on mibe-roberta-ff-bilstm. *Scientific Reports*, 14(1):5827, 2024. 2

[45] H. Zuo, O. Wu, W. Hu, and B. Xu. Recognition of blue movies by fusion of audio and video. In *2008 IEEE International Conference on Multimedia and Expo*, page 37–40, Jun 2008. 3