Labeling Comic Mischief Content in Online Videos with a Multimodal Hierarchical-Cross-Attention Model

Elaheh Baharlouei[1], **Mahsa Shafaei**[1], **Yigeng Zhang**[1]
Hugo Jair Escalante[2], **Thamar Solorio**[1,3]
[1] Department of Computer Science, University of Houston, Houston, USA
{ebaharlouei, mshafaei, yzhang168, tsolorio}@uh.edu
[2] Department of Computer Science, INAOE, Puebla, Mexico
hugojair@inaoep.mx
[3] MBZUAI, Masdar City, Abu Dhabi, UAE

Abstract

We address the challenge of detecting questionable content in online media, specifically the subcategory of comic mischief. This type of content combines elements such as violence, adult content, or sarcasm with humor, making it difficult to detect. Employing a multimodal approach is vital to capture the subtle details inherent in comic mischief content. To tackle this problem, we propose a novel end-to-end multimodal system for the task of comic mischief detection. As part of this contribution, we release a novel dataset for the targeted task consisting of three modalities: video, text (video captions and subtitles), and audio. We also design a HIerarchical Cross-attention model with CAPtions (HICCAP) to capture the intricate relationships among these modalities. The results show that the proposed approach makes a significant improvement over robust baselines and state-of-the-art models for comic mischief detection and its type classification. This emphasizes the potential of our system to empower users, to make informed decisions about the online content they choose to see. In addition, we conduct experiments on the UCF101, HMDB51, and XD-Violence datasets, comparing our model against other state-of-the-art approaches showcasing the outstanding performance of our proposed model in various scenarios.

Keywords: Multimedia Document Processing, Social Media Processing, Tools, Systems, Applications

1. Introduction

The impact of media on children has long been debated in psychology (Dietz and Strasburger, 1991). Regardless of the positive impacts of media on children's education, multiple studies demonstrate that violent and aggressive material negatively affects children's behavior (Wilson, 2008; Chang and Bushman, 2019). Online media may have detrimental effects beyond violence, for example, by normalizing negative behaviors such as drug and alcohol abuse or premature engagement in sexual activities in society (Hanewinkel et al., 2014; Strasburger, 1989).

Automated systems for categorizing online content based on the presence of questionable material have the potential to make a significant difference in protecting users from distressing and possibly objectionable content. Identifying content with intuitive labels offers a more flexible alternative to the commonly used age-based rating systems, such as those by the Motion Picture Association of America. This flexibility accounts for the fact that people's tolerance for questionable content can vary widely, influenced by factors like age, life experiences, socio-cultural values, and cognitive skills (Anderson et al., 2003). In this paper, for the first time, we focus on detecting comic mischief content in videos, which is a subset of questionable content. In a comic mischief video, questionable content (violence, adult content, or sarcastic material) is combined with a humorous context, making it even more disruptive. According to psychologists, when something such as violence is presented in a serious context (such as war), it has a less disruptive effect than when it is presented in a pleasant and humorous context (Blackford et al., 2011).

Automated detection of comic mischief in videos poses a significant challenge. The difficulty lies in differentiating between serious content and humor, particularly when the humor involves subtle jokes or culturally specific knowledge. People's perceptions of humor can vary greatly due to cultural backgrounds and personal preferences, complicating the establishment of universal criteria for humor detection. Moreover, videos comprise visual imagery, audio, and text, necessitating a comprehensive approach that considers all these elements (Yang et al., 2023b). In response to these challenges, our paper introduces a multimodal approach designed to capture the diverse information presented among the modalities to better understand and identify comic mischief. We present HICCAP, an innovative multimodal system specifically developed for detecting comic mischief. It features a hierarchical cross-attention module that identifies intermodal relationships, enhancing our system's effectiveness. Our approach includes a binary model for determining the presence of comic mischief and a multi-task model that categorizes different types of comic mischief. The

LREC-COLING 2024, pages 9999–10013

models analyze three key modalities found in our dataset: visual content, audio signals, and textual information from dialogues. The proposed model is pretrained using a multimodal hybrid pertaining technique that integrates both matching tasks and contrastive-learning methods on extensive multimodal datasets. Furthermore, our empirical research demonstrates that using a pretrained video captioning model effectively supplements subtitles in videos that lack them and boosts the model's performance.

Complementing our multimodal system, we introduce a new dataset for comic mischief task, collected from a mix of freely accessible internet videos and the Youtube-8M (Abu-El-Haija et al., 2016) corpus. This collection has been meticulously annotated for comic mischief categories by twelve annotators. This dataset can be a valuable resource for the community by giving detailed information about the comic mischief content across multiple video modalities (i.e., dialogue, sound, and video). While it will chiefly benefit those in natural language processing, computer vision, machine learning, and signal processing, the dataset also holds the potential to inspire new inquiries in social sciences, child psychology, and mass media fields.

The contributions of this paper are as follows:

- Introduces a new video classification task and provides a novel dataset aiming to motivate further research on this relevant topic[1].
- Proposes a new end-to-end multimodal model based on a hierarchical cross-attention module to effectively capture the relationships between different modalities.
- Incorporates an automatic captioning technique to fill up the gaps in video subtitles that considerably boosts the model's performance.

2. Multimodal Comic Mischief Dataset

To create technology for labeling comic mischief in videos, we need access to a repository of consistently labeled video content. This requires collecting a diverse selection of online videos and using a customized annotation tool to implement an incremental annotation strategy. In this section, we introduce the comic mischief dataset we are releasing with this paper.

In our dataset, each comic mischief label may be associated with any or all three considered modalities (sound, dialogue, and video). The definition of each modality is as follows: **Dialogue:** transcription of spoken dialogue between characters (subtitle of video); **Sound:** sound effects and ambient sounds (e.g., explosions); **Video:** intensity (the amount of light or the numerical value of a pixel) information from video frames. We considered four categories of comic mischief including:

- **Gory Humor:** refers to a situation with a great deal of bloodshed and violence juxtaposed with humorous references.
- **Slapstick Humor:** is a comedy style characterized by practical jokes, collisions, clumsiness, and embarrassing events (e.g., people get poked in the eye or pies in the face).
- **Mature Humor:** strong language, alcohol/drug consumption, gambling, and sexual references in depictions or dialogue with humor references.
- **Sarcasm:** employs words to mock or annoy someone for humorous effect. Sarcasm may employ ambiguity, but it is not always ironic.

Figure 1 shows sample screenshots of comic mischief categories.

(a) Gory Humor (b) Slapstick

(c) Mature Humor (d) Sarcasm

Figure 1: Comic mischief examples in movies

2.1. Video Collection Process

Our goal is to collect a wide range of English-language videos with comic mischief material. For our data collection, we leverage two primary sources. Firstly, we gather data from YouTube, ensuring that our methods adhere to YouTube's policies. In alignment with these policies, we will release a JSON file containing relevant YouTube URLs for the dataset, making it accessible for viewing and downloading. YouTube is one of our platforms of choice given its widespread availability to internet users. We developed an algorithm that utilizes YouTube's recommendation system to discover new videos. The methodology includes: 1) Conducting an initial manual search to identify seed videos on the desired topic. 2) Generating a list of videos recommended for each seed video. Given YouTube's content-based recommendation mechanism, some content overlap is anticipated. These suggested videos then became our new seed material. 3) Repeating the above process three times to ensure diversity among videos. To

[1] https://github.com/RiTUAL-UH/Comic-Mischief-Prediction

maintain consistency in theme, with each iteration, we selected fewer videos per seed. Our second resource is the YouTube-8M dataset, which is a freely available dataset for the research community. This corpus is a collection of human-verified labels on about 237K segments on 1000 classes. To ensure varied video content, our primary focus is on categories such as talk shows, sketch comedy, comedy (drama), sitcoms, and stand-up comedies.

To facilitate our research, we divided the videos into 60-second clips and proceeded to annotate these clips. In the remainder of this paper, we use the term clip while referring to the 60-second segments of a video in our dataset. In total, we have 4478 clips extracted from 1179 videos. From the first source, YouTube, we collected 1233 clips originating from 394 videos. From the YouTube-8M dataset, our second source, we gathered 3245 clips derived from 785 videos.

2.2. Annotation Process

As a first step, we must enrich the taxonomy of questionable content and establish guidelines for annotators. To create a guideline for annotation, we did the following steps: 1) established criteria for annotation of questionable content and developed annotation guidelines; 2) compiled a list of diverse online videos for use in pilot annotations; 3) to revise the annotation process and components, we conducted several workshops with participants from the fields of psychology, AI ethics, computer vision, and natural language processing; 4) finally, we fine-tuned the guidelines in light of pilot annotation experience and discussion. To ensure the quality of the annotation process, we have implemented a three-way web-based annotation interface displaying each clip with its audio and transcript. This interface facilitates annotations across various comic mischief categories in three modalities, with each clip being evaluated by three annotators. Figure 2 illustrates our web interface. Twelve annotators participated in the task, including seven graduate and five bachelor students. Of these, seven were native English speakers, while the remaining five were fluent in English. Each label has a binary value, indicating the presence of comic mischief. Also, each label can be displayed in any of the clip's modalities (Dialogue, Sound, Video). As a result, we have distinct labels for each modality (e.g., Gory Humor - Dialogue, Gory Humor - Video) and one clip can belong to more than one category. A majority vote of all annotators determines the final value assigned to each label per modality. To evaluate the annotation's quality, we have computed the Inter-Annotator Agreement (IAA) by calculating Cohen's Kappa (κ) on each annotator's annotation and the majority voting of all annotations. Based on the computed IAA values, there is a substantial agreement ($\kappa = 0.70$).

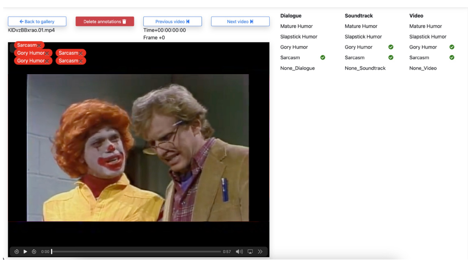


Figure 2: User interface for labeling dataset videos

2.3. Dataset Statistics

Table 1 provides an overview of the class-level statistics for the video segments. As the table shows, some videos do not contain dialogue (e.g., silent films). Videos in class 1 (C1) contain at least one type of comic mischief, while those in class 0 (C0) do not. This table shows that each class has an equal distribution of video length and content. The statistics demonstrate that the length of videos and the average number of words per video are comparable for both classes, ensuring that they contain enough content to be meaningful.

	Max		Min		Avg		Med	
	C0	C1	C0	C1	C0	C1	C0	C1
# Words	259	266	0	0	106	118	111	125
V/A Length	64.9	71.9	0.1	9.4	54.7	58.6	60.1	60.5
# Frames	1836	2157	1	108	538	658	460	478

Table 1: Video segments statistics, C0, C1, and V/A Length stand for class 0, class 1, and video or audio clip length respectively.

Figure 3 displays the data distribution across each category for the multi-label dataset. As shown in this figure, classes are not balanced: the majority class (Mature Humor) has more than five times the samples for the minority class (Gory humor).

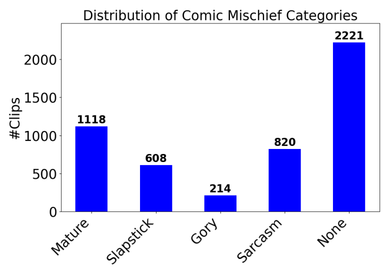


Figure 3: Distribution of Comic mischief categories.

The introduced dataset is the first one approaching the comic mischief detection task. This dataset poses a number of challenges including: video classification with imbalanced classes, the need for effective use of multimodal information for classification, understanding subtle humor cues intertwined with potentially harmful content, discerning the intricate relationships between visual, audio, and textual modalities in the context of comic mischief, and handling data where audio and text may sometimes conflict or diverge from visual content.

3. Methodology

In this section, we describe an automated model for categorizing videos containing comic mischief material. We approach this problem from two perspectives: first, as a binary classification task, and second, by developing a multi-task model that can concurrently predict all four subcategories.

Our "HIerarchical Cross attention model with CAPtions" (HICCAP) is a unified end-to-end model designed to capture the inherent complexities of comic mischief prediction and type classification. We encode each modality (audio, text, and video) with pretrained models, and then fed the encoded audio and video vectors into RNNs to account for the sequential information. Recognizing that one modality can impact the interpretation or significance of others (e.g., the speaker's tone may alter the meaning of their words), we introduce a cross-attention network to capture these intermodality relationships. Figure 4 provides the system overview, with subsequent sections detailing each module.

3.1. Feature Encoding

To better represent a video, we utilize pretrained models to extract features from various modalities, including raw text, audio, and video, ensuring a comprehensive understanding of the context.

We use the pretrained BERT model (Devlin et al., 2018) for encoding the input text in our model. Since some of the videos lack dialogue, they do not include subtitles. To fill this void, we use a pre-trained model for video captioning to generate captions for these videos and use them in the absence of subtitles. We employ Dense video Captioning with a Bi-modal Transformer (BMT), for this purpose (Iashin and Rahtu, 2020). In BMT, the model locates important events in a video and generates a unique textual description for each event.

We utilize the pretrained VGGish network (Hershey et al., 2017) for audio feature extraction and the pretrained I3D network (Carreira and Zisserman, 2017) for visual modality encoding. For both audio and video vectors, we process them through an LSTM to extract sequential information, followed by a Fully Connected (FC) layer.

3.2. Hierarchical Cross-Attention

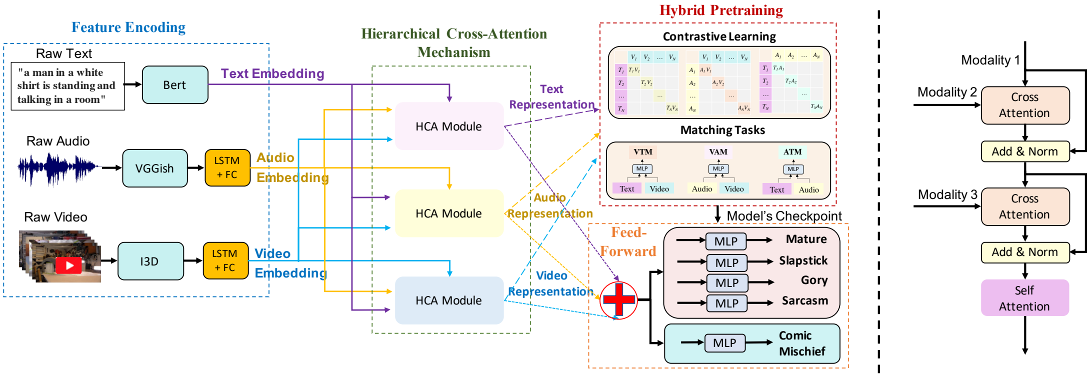
In multimedia videos, the content of each modality may affect the meaning or importance of other modalities. To capture this intermodality relation, we utilize the cross-attention mechanism to enhance the representation of multiple modalities. In cross-attention, the query belongs to one modality, while the key and value vectors belong to the context modality (Zheng et al., 2020; Lu et al., 2019). To leverage the effectiveness of cross-attention for multiple modalities we introduce a new hierarchical cross-attention mechanism.

While the concept of hierarchical cross attention has been explored in earlier studies (Chen et al., 2022; Zhang et al., 2022; Yang et al., 2023a), our proposed Hierarchical Cross Attention technique, HICCAP, offers unique features and benefits. Notably, the methods presented in (Chen et al., 2022; Yang et al., 2023a) are restricted to a single modality, and (Zhang et al., 2022) focuses solely on image and text. In contrast, HICCAP encompasses three modalities: video, audio, and text, and uniquely aims to encode the significance of a single modality based on the interactions with the other modalities. To our knowledge, it is the first work that considers all these modalities together. Additionally, HICCAP introduces a distinct hierarchical cross-attention (HCA) module (Figure 4.b) for every modality and strategically combines these HCA modules, ensuring that while emphasizing one modality, the attention from the other two modalities is also considered.

In our hierarchical cross-attention approach, the first cross-attention layer calculates the attention for modality M_1 using M_2. This output subsequently forms the query vector, while modality M_3 serves as both key and value vectors for the following cross-attention layer. This ensures that M_1's attention is influenced by both M_2 and M_3. The procedure can be formalized as shown in Equation 1, where K, Q, and V denote key, query, and value respectively, and d_k represents the key vector's dimension. In the hierarchical cross-attention approach, the order in which modalities are processed is pivotal. The sequence dictates how information flows and integrates, influencing the final representation. Consequently, selecting an appropriate ordering for modalities, based on the characteristics of the dataset and task, can impact the model's effectiveness and understanding. We did an empirical experiment to find the best ordering of modalities in different hierarchical attention modules.

$$
\begin{aligned}
head_1 &= \text{softmax}(\frac{K_{m2}^T Q_{m1}}{\sqrt{d_k}})V_{m2} \\
head^{m1} &= \text{softmax}(\frac{K_{m3}^T Q_{head_1}}{\sqrt{d_k}})V_{m3}
\end{aligned}
\tag{1}
$$

In Equation 1, $head^{m1}$ shows the representation of modality M_1 based on modalities M_2 and M_3. We repeat this process for calculating $head^{m2}$ and $head^{m3}$ which are the representations



a) The general architecture of HICCAP

b) HCA Module

Figure 4: a) The general architecture of HICCAP consists of four components: 1) Feature-Encoding, 2) Hierarchical-Cross-Attention mechanisms, 3) Pretraining, and 4) Binary and Multi-Task Prediction and b) the structure of Hierarchical Cross Attention (HCA) Module.

of M_2 and M_3, respectively. Then, we pass each representation separately through the attention layer (Bahdanau et al., 2014). This layer computes the weighted sum r as $\sum_i \alpha_i head_i^{mj}$ where $j \in \{1, 2, 3\}$ to aggregate hidden layers of the cross-attention layer to a single vector. The model can learn the relative importance of hidden states ($head_i^{mj}$) by learning the α_i. We compute α_i as: $\alpha_i = \text{softmax}(v^T \tanh(W_h head_i^{mj} + b_h))$ where W_h is the weight matrix, and v and b_h are the parameters of the network.

3.3. Pretraining Approach

We leverage the Video-Text Matching (VTM) pretraining task and also adopt analogous approaches for Video-Audio Matching (VAM) and Audio-Text Matching (ATM) to enhance the model's performance. Alongside this, we integrate contrastive learning during pretraining to enhance multimodal representation learning. Combining VTM, VAM, and ATM tasks with contrastive learning establishes a robust pretraining framework, aligning various modalities in a unified representation space. This joint pretraining approach enables the model to capture intricate relationships between video, audio, and text which leads to improved performance on downstream tasks. Following pretraining, we fine-tune the pretrained model for comic mischief detection and subtype classification. More details about pretraining approaches have been provided in Appendix (9.1).

3.4. Multi-task learning

It is well known that multi-task learning can improve performance when there is shared information between tasks, as seen in similar studies (Zhang et al., 2021, 2023b). Although in our problem we only have a single task, the considered categories may be correlated as all tasks care about the comic aspect accompanied by a kind of objectionable content. Therefore we adopted a multi-task learning setting for training the model in the multi-label classification scenario. We added separate MLP blocks for each class as can be seen in Figure 4, part 4. In multi-task learning, each task will have its unique loss function, denoted by L_i. We weight each loss function and minimize a linear combination of these weighted losses; weights are learnable parameters that will be tuned during training $L_{total} = \sum_i w_i L_i$.

4. Experimental evaluation

4.1. Experimental Setup

We utilize a subset (50K) of the HowTo100M (Miech et al., 2019) dataset for matching (VTM, VAM, and IAM) pretraining tasks and Kinetics-400 (Kay et al., 2017) for contrastive learning purposes in compliance with their policies.

Data Partitions: We divided the dataset into three parts: 65% for training, 10% for validation, and 25% for testing. Table 2 shows the data distribution of the dataset for each partition.

	MH	SH	GH	S	None	# Clips
Train	784	430	152	578	1297	2890
Validation	99	93	29	94	169	432
Test	222	116	40	158	267	1156

Table 2: Dataset partitions. Columns 2-4 indicate number of samples per category; column 6 shows the total number of clips per partition. MH, SH, GH, and S stand for mature humor, slapstick humor, gory humor, and sarcasm, respectively. Note that one clip can belong to more than one category.

Metrics: We report F1 score with respect to the positive class for the binary model and macro F1 score for multi-task model.

4.2. Ablation and Analysis

In this section, we carry out a comprehensive set of experiments to assess the performance of our proposed model in comparison to a range of baseline models, focusing on two tasks: first, the binary classification of comic mischief presence, and second, the multi-task prediction of specific comic mischief subcategories. In our evaluation, we have selected several baseline systems for comparison. The baseline categories include 1) unimodal systems , 2) bi-modal approaches, and 3) models that consider all three modalities but do not employ a hierarchical cross-attention mechanism.The rationale behind selecting these baselines is to demonstrate the importance of both multi-modal integration and the hierarchical cross-attention mechanism for accurately predicting comic mischief content and its type. Additionally, we conducted experiments to highlight the significance of pretraining techniques.

Impact of Multimodality: We provide a comparative analysis between the single modality model, the Bi-modal model, and our HICCAP for binary and multi-task prediction. In single modality models, we employ a pretrained BERT with attention and fully connected (FC) layers for text, and deploy LSTM integrated with attention and FC layers for both audio and video modalities. The Bi-modal models integrate the original cross-attention mechanism limited to two modalities. Our results in Table 3.a indicate that HICCAP significantly improves upon the best-performing monomodal and Bi-modal cross-attention models by 11.12% and 8.06%, respectively, in terms of F1 score, for the binary classification task. Also in multi-task experiments, HICCAP outperforms these models by 9.8% and 6.94%, respectively, based on the average macro F1. The improvements are statistically significant with a p-value < 0.001 according to the Mcnemar significance test.

Impact of Hierarchical Cross Attention: To assess the contribution of the hierarchical cross attention, we compare HICCAP against multimodal concatenation models including early and late fusion techniques. The early fusion strategy integrates modalities by concatenating feature vectors, while the late fusion trains each modality independently and combines them at the decision level by averaging class probabilities. Table 3.b showcases that HICCAP outperforms both early and late fusion models, by 7.79% and 7.03%, respectively, based on the F1 score for binary classification. Besides, in this table for multi-task prediction, HICCAP improves the performance of early and late fusion models by 7.83% and 7.09%, respectively, based on macro average F1 scores.

Impact of Adapting Caption Generation: To illustrate the influence of captioning on videos without subtitles, we incorporated captions into both the text single modality and hierarchical cross-attention (HCA) models for binary and multi-task prediction. The results in Table 3.c reveal improvements for both models, with HICCAP surpassing the HCA model by 1.36% in terms of F1 score, for binary classification. Also, the results show applying the caption generation technique improves the performance of the HCA model by 0.39% for multi-task prediction. However, adding captions is not as effective as in the binary model.

Impact of Pretraining Techniques: To show the importance of pretraining, we conduct experiments focusing on matching pretraining only, contrastive learning (CL) only, and a combination of both. The data in Table 4 highlights that HICCAP with hybrid pretraining surpasses both the non-pretrained variant and those with matching and CL pretraining, showing gains of 4.18%, 2.72%, and 1.84%, respectively, base on F1 score, for binary classification. Furthermore, the findings in this table demonstrate that for multi-task prediction employing a hybrid pretraining method leads to a significant improvement in HICCAP performance, in terms of average macro F1 score.

Impact of Multi-Task Learning: Table 5 compares the result of the multi-task setting of our model with the four single-task hierarchical cross-attention models. All the models are exactly the same, but we trained them for each task separately. The multi-task model works better overall based on the average macro F1. Moreover, to show a hierarchical cross-attention mechanism in the single-task models is a reasonable approach, we compare the single-task models with the late fusion and intermediate GMU fusion models. The hierarchical cross-attention model outperforms late fusion and GMU fusion by 10.12% and 22.77%, respectively, based on macro F1. Also, as shown in this table, the multi-task approach performs better than the multi-label by 1.86% in terms of macro F1. Given the superior performance of the multi-task strategy, we decided to adopt this approach for the architecture of our proposed HICCAP model. This choice was motivated by our aim to maximize the effectiveness and efficiency of the model in addressing the complexities of the underlying tasks.

Model Evaluation: We consider the following multimodal methods to evaluate the binary model:

LXMERT (Tan and Bansal, 2019): To make this framework compatible with our work, we pass the video-encoded vector (I3D features) to the model rather than object vectors.

X-CLIP(Ni et al., 2022): In order to incorporate X-CLIP with our structure, we utilized the checkpoint provided by the authors for zero-shot training. Subsequently, we fine-tuned and assessed the model

Models	Modality	BM F1-Score	MTM Macro F1
Single Modal	T(+C)	60.60	60.15
	A	60.08	59.9
	V	59.93	58.27
Bi Modal	T(+C)+V	62.79	62.35
	A+V	63.66	63.01
	T(+C)+A	63.12	62.69
HICCAP	T(+C)+ A+V	**71.72**	**69.95**

a) Multimodality

Models	Modality	BM F1-Score	MTM Macro F1
Early Fusion	T(+C)+ A+V	63.93	62.12
Late Fusion	T(+C)+ A+V	64.69	62.86
HICCAP	T(+C)+ A+V	**71.72**	**69.95**

b) Hierarchical Cross Attention

Models	Modality	BM F1-Score	MTM Macro F1
Single Modal	T	59.12	58.85
	T(+C)	60.60	60.15
HCA	T+A+V	70.36	69.56
HICCAP	T(+C)+ A+V	**71.72**	**69.95**

c) Adapting caption generation

Table 3: Ablation study on the effect of a) Multimodality, b) Hierarchical Cross Attention, and c) Adapting caption generation, for comic mischief detection in binary model (BM) (Column 3) and type prediction in multi-task model (MTM)(Column 4). 'T', 'A', 'V', and 'C' stand for text, audio, video, and caption.

Models	Modality	BM F1-Score	MTM Macro F1
w/o Pretrain	T(+C)+ A+V	71.72	69.95
Matching	T(+C)+ A+V	72.24	72.48
CL	T(+C)+ A+V	73.12	73.30
Hybrid CL +Matching	T(+C)+ A+V	**74.96**	**74.13**

Table 4: Ablation study on the effect of pretraining techniques for comic mischief detection and type classification.

Method	F1 MH	F1 GH	F1 SH	F1 S	Macro F1
Late fusion per task	67.05	42.08	51.83	69.83	57.69
GMU fusion per task	49.14	46.73	45.93	38.37	45.04
HICCAP per task	79.44	73.59	43.80	74.42	67.81
Multi-label HICCAP	78.64	57.54	77.36	58.83	68.09
Multi-task HICCAP	76.83	65.73	62.34	74.93	**69.95**

Table 5: Comparing a) multi-task model with multi-label and single-task models, b) HCA per task with late and GMU fusion.

Models	Modality	F1-score
GMU	T(+C)+ A+V	46.90
LXMERT	T(+C)+V	64.24
	T(+C)+A	63.61
X-Clip	T(+C)+V	68.80
HICCAP	T(+C)+ A+V	71.72
Pretrain HICCAP	T(+C)+ A+V	**74.96**

Table 6: Comparison of HICCAP with previous methods for binary comic mischief detection.

using the comic mischief dataset.

Intermediate fusion with GMU: As with the MM-trailer model (Shafaei et al., 2021), we train the model for each modality separately, then save the final layer before the classification layer, and ultimately use the GMU model to integrate the output from all modalities.

Table 6 shows the performance of the binary model for comic mischief detection. The table shows that the best results were obtained from the HICCAP with the hybrid pertaining technique. To fairly evaluate the HICCAP mechanism and compare it to other methods, we also report results for models without pretraining. The results show that HICCAP improves the best intermediate GMU fusion (Shafaei et al., 2021), LXMERT (Tan and Bansal, 2019), and X-CLIP (Ni et al., 2022) models by 24.82%, 7.48%, and 2.92%, respectively, based on F1 score.

4.3. Comparison with State-of-the-Art

We assess the performance of our HICCAP on two multimodal datasets designed for activity recognition: UCF-101 (Soomro et al., 2012) and HMDB51 (Kuehne et al., 2011), as well as on the XD-Violence (Wu et al., 2020) dataset tailored for anomaly detection. For caption generation on these datasets, we employ the prompt generation technique presented in X-Clip (Ni et al., 2022). In this method, the authors enhanced the original text encoder, initially trained for language-image tasks, with a video-specific prompting mechanism. By utilizing video content features, they aimed to optimize the text prompting process, emphasizing that proper contextual information boosts recognition.

4.3.1. Result on UCF101 and HMDB51

Table 7 summarizes the performance comparison between the proposed method and other state-of-the-art methods on the UCF101 and HMDB51 datasets. In the presented table, our method outperforms the majority of previously established techniques in terms of top-1 accuracy. Notably, there is a single method, VideoMAE V2 (Wang et al., 2023) that our proposal does not outperform. However, it's crucial to highlight that even in this case, our performance is comparable to this paper. Importantly, our method achieves this with a more straightforward strategy and a model with significantly fewer parameters. This underscores the efficacy and efficiency of our approach, demonstrating that competitive outcomes can be realized without the need for complex systems or excessive model parameters.

4.3.2. Result on XD-Violence

In this section, we applied our method to the XD-Violence dataset. Due to the lack of access to part of the provided vector features, we regenerate the I3D and VGGish features using the provided training and test videos. Similar to (Wu and Liu, 2021), we use the frame-level precision-recall curve (PRC) and corresponding area under the curve (average precision, AP) and report the results in Table 8. As shown in this table, our method achieves the new

Method	Modality	UCF101	HMDB51
VideoMoCo (Pan et al., 2021)	V	78.7	49.2
Vi2CLR (Diba et al., 2021)	V	89.1	55.7
CVRL (Qian et al., 2021)	V	94.4	70.6
CORPf (Hu et al., 2021)	V	93.5	68.0
MIL-NCE (Miech et al., 2020)	T+V	91.3	61.0
MMV (Alayrac et al., 2020)	T+A+V	92.5	69.6
CPD (Li and Wang, 2020)	T+V	92.8	63.8
ELO (Piergiovanni et al., 2020)	A+V	93.8	67.4
XDC (Alwassel et al., 2020)	A+V	94.2	67.1
GDT (Patrick et al., 2020)	A+V	95.2	72.8
VideoMAE V1(Tong et al., 2022)	V	96.1	73.3
VideoMAE V2(Wang et al., 2023)	V	99.6	88.1
Ours	T+A+V	**98.87**	**76.64**

Table 7: Comparison with UCF101 and HMDB51.'T', 'A', 'V', and 'C' stand for text, audio, video, and caption.

state-of-the-art performance of 92.17% AP and gains clear improvements when compared with previous SOTA methods in terms of AP.

Method	Modality	AP (%)
AVVD (Wu et al., 2022b)	V	78.10
NG-MIL (Park et al., 2023)	V	78.51
MGFN (Chen et al., 2023)	V	80.11
SR3 (Wu et al., 2022a)	V	80.26
CU-Net (Zhang et al., 2023a)	A+V	81.43
DMU (Zhou et al., 2023)	A+V	81.77
CLIP-TSA (Joo et al., 2023)	V	82.19
MACIL-SD (Yu et al., 2022)	A+V	83.4
DDI (Pu and Wu, 2022)	A+V	83.54
Song et al. (Fan et al., 2023)	A+V	84.23
VadCLIP (Wu et al., 2023)	V+T	84.51
Ours	T+A+V	**92.17**

Table 8: Comparison with SOTA on XD-Violence. 'T', 'A', 'V', and 'C' stand for text, audio, video, and caption.

4.4. Error Analysis

In the analysis of misclassified instances of comic mischief, approximately 80% belong to a single category (such as sarcasm exclusively), while the remainder involved multiple categories. This observation aligns with the anticipation that videos with multiple categories tend to be more accurately identified as instances of comic mischief. Also, within the "none" category, about 35% of predictions were incorrectly predicted. This could be due to the presence of mature or violent content lacking humorous elements, or possibly from the smaller number of "none" examples, complicating the model's ability to discern a clear pattern.

Our investigation into the reliance of the model on various modalities revealed enlightening insights, particularly with respect to sarcasm—a category predominantly inferred from textual dialogues. An experimental manipulation involved masking the video and audio components for clips categorized under sarcasm, leading to mispredictions by our model. This outcome underscores a critical observation: despite sarcasm's heavy reliance on textual cues within dialogues, the integration of video and audio modalities plays a non-negligible role in enhancing the model's predictive performance. Extending this line of inquiry to the gory and slapstick humor categories, which are primarily video-centric, we conducted similar experiments by masking text and audio modalities to observe the impact on multimodal prediction capabilities. The modifications led to a significant degradation in the model's performance for these categories. These mispredictions highlight the indispensable value of multimodal data fusion in understanding complex comic mischief categories, thereby emphasizing the need for a comprehensive approach in the processing and interpretation of multimodal information for accurate content categorization.

5. Related Work

Multimodal Machine Learning (MML). MML (Baltrušaitis et al., 2018; Xu et al., 2022) has gained considerable attention in recent decades as a vital research area. In recent literature, numerous models have been proposed to analyze multimodal information in videos, taking into account the complex interplay between visual, auditory, and textual data (Seo et al., 2022; Man et al., 2022; Basak et al., 2022; Sun et al., 2022). In (Akbari et al., 2021), the authors present a Video-Audio-Text Transformer (VATT) for learning multimodal representations from unlabeled data using convolution-free transformer architectures. VATLM (Zhu et al., 2022) combines video, audio, and text modalities by utilizing a unified transformer tokenizer following modality-specific encoders and subsequently performing masked prediction on the integrated tokens. Compared with previous work, HICCAP introduces a novel method that combines various Hierarchical multimodal Cross-Attention (HCA) mechanisms. Each HCA learns cross-attention between one modality and the other two, producing unique representations for each modality, ultimately leading to improved model performance.

Cross-Modal Pretraining Tasks. Several studies have focused on pretraining tasks to warm-start multimodal model parameters, aiming to boost performance in downstream tasks. One popular pretraining task is Image-Text Matching (ITM) which aims at grasping the coarse-grained correlation between images and texts (Huang et al., 2021; Li et al., 2020; Du et al., 2022). Recent works like (Li et al., 2021) have employed a multimodal encoder with cross-attention for the ITM task. Contrastive learning (CL) has also seen significant advancements in representation learning, both in unimodal (Chen et al., 2020; Gao et al., 2021; Oord et al., 2018) and multimodal contexts (Li et al., 2022a, 2023; Yuan et al., 2021; Ramesh et al., 2021; Udandarao et al., 2020; Goel et al., 2022). Notable implementations, such as CLIP (Radford et al., 2021), align text-image pairs, while its extension, X-Clip (Ni et al., 2022), focuses on video recognition. LAVA (Gurram et al.) delves into video-audio-text multimodal pretraining. Recent advancements in multimodal learning have highlighted the benefits of hybrid pretraining techniques. Notably, BLIP(Li et al., 2022a) introduces a combination of image-text matching and image-text contrastive pretraining. Further, BLIP-2 (Li et al., 2023) expands the approach by jointly optimizing three objectives: Image-Text Matching, Image-Text Contrastive Learning, and Image-Grounded Text Generation. In this paper, HICCAP employs a multimodal video-audio-text CL strategy, combined with video-audio-text matching pretraining techniques, aiming for the integrated learning of language, audio, and video representations.

6. Conclusion

We introduced the task of labeling comic mischief in videos and released a dataset for this purpose. We also proposed a new model named "HIerarchical Cross attention model with CAPtions" (HICCAP). HICCAP uses hierarchical cross-attention modules to capture interactions between the three modalities. We pretrained the HICCAP model with hybrid modality matching and contrastive learning before fine-tuning it for binary and multi-task classifications. The experiments indicate that our method outperforms both the baseline and other reference models. Our main conclusions are the following:

- *Detecting comic mischief in videos is a feasible task for multimodal learning.* Likewise, there is still too much room for improvement in this relevant task, therefore, we foresee it will be of interest to the community.
- *Hierarchical cross-attention (HCA) obtained better performance than standard cross-attention.* Contrary to most work on multimodal learning, HCA is able to capture dependencies across all of the considered modalities, this way of modeling dependencies resulted in better performance when compared to models that only implemented self and bimodal cross-attention.
- *Both, multimodal matching pretraining and contrastive learning pretraining proved to be useful to warm-start the model, and their combination further boosted the model's performance.* We show that the adoption of the hybrid pretraining process resulted in the best performance for our HICCAP model.

Future work includes the exploration of hierarchical cross-attention in the context of other multimodal learning models. Also, we would like to develop explainable models for detecting comic mischief content.

7. Acknowledgments

We're grateful to the anonymous reviewers for their valuable comments. We appreciate the support from Ioannis Kakadiaris and Christos Smailis during the dataset annotation process. We also thank Jared Suchomel, Alli Brophy, Lyndon Lam, Connor Maguire, and Kelly Couvrette for their assistance with dataset annotation. This research was partially funded by the National Science Foundation, under awards 1910192 and 2106892.

8. Bibliographical References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221.

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37.

Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770.

Craig A Anderson, Leonard Berkowitz, Edward Donnerstein, L Rowell Huesmann, James D Johnson, Daniel Linz, Neil M Malamuth, and Ellen Wartella. 2003. The influence of media violence on youth. *Psychological science in the public interest*, 4(3):81–110.

Dzmitry Bahdanau, Kyunghyun Cho, et al. 2014. Neural machine translation by jointly learning to align and translate. arxiv preprint arxiv: 1409.0473.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Hritam Basak, Rohit Kundu, Pawan Kumar Singh, Muhammad Fazal Ijaz, Marcin Woźniak, and Ram Sarkar. 2022. A union of deep learning and swarm-based optimization for 3d human action recognition. *Scientific Reports*, 12(1):5494.

Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. 2020. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931.

Benjamin J Blackford, James Gentry, Robert L Harrison, and Les Carlson. 2011. The prevalence and influence of the combination of humor and violence in super bowl commercials. *Journal of Advertising*, 40(4):123–134.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Justin H Chang and Brad J Bushman. 2019. Effect of exposure to gun violence in video games on children's dangerous behavior with real guns: a randomized clinical trial. *JAMA network open*, 2(5):e194319–e194319.

Shuning Chang, Yanchao Li, Shengmei Shen, Jiashi Feng, and Zhiying Zhou. 2021. Contrastive attention for video anomaly detection. *IEEE Transactions on Multimedia*, 24:4067–4076.

Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. 2021a. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8012–8021.

Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. 2021b. Rspnet: Relative speed perception for unsupervised video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1045–1053.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xin Chen, Ben Kang, Dong Wang, Dongdong Li, and Huchuan Lu. 2022. Efficient visual tracking via hierarchical cross-attention transformer. In *European Conference on Computer Vision*, pages 461–477. Springer.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. 2023. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 387–395.

J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. 2021. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1502–1512.

William H Dietz and Victor C Strasburger. 1991. Children, adolescents, and television. *Current problems in pediatrics*, 21(1):8–31.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*.

Yidan Fan, Yongxin Yu, Wenhuan Lu, and Yahong Han. 2023. Weakly-supervised video anomaly detection with snippet anomalous attention. *arXiv preprint arXiv:2309.16309*.

Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. 2021. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. 2022. Cyclip: Cyclic contrastive language-image pre-training. *Advances in Neural Information Processing Systems*, 35:6704–6719.

Sumanth Gurram, David Chan, Andy Fang, and John Canny. Lava: Language audio vision alignment for data-efficient video pre-training. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Tengda Han, Weidi Xie, and Andrew Zisserman. 2020a. Memory-augmented dense predictive coding for video representation learning. In *European conference on computer vision*, pages 312–329. Springer.

Tengda Han, Weidi Xie, and Andrew Zisserman. 2020b. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690.

R. Hanewinkel, J. D. Sargent, K. Hunt, H. Sweeting, R. C. Engels, R. H. Scholte, E. Mathis, F.and Florek, and M. Morgenstern. 2014. Portrayal of alcohol consumption in movies and drinking initiation in low-risk adolescents. *Pediatrics*, 133(6):973–982.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. 2021. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7939–7949.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985.

Vladimir Iashin and Esa Rahtu. 2020. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271*.

Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. 2023. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3230–3234. IEEE.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE.

Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2017. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 667–676.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Shuo Li, Fang Liu, and Licheng Jiao. 2022b. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1395–1403.

Tianhao Li and Limin Wang. 2020. Learning spatiotemporal features via video and text pair discrimination. *arXiv preprint arXiv:2001.05691*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Xin Man, Deqiang Ouyang, Xiangpeng Li, Jingkuan Song, and Jie Shao. 2022. Scenario-aware recurrent transformer for goal-directed video captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(4):1–17.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 1–18. Springer.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11205–11214.

Wen-Feng Pang, Qian-Hua He, Yong-jian Hu, and Yan-Xiong Li. 2021. Violence detection in videos based on fusing visual and audio information. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2260–2264. IEEE.

Seongheon Park, Hanjae Kim, Minsu Kim, Dahye Kim, and Kwanghoon Sohn. 2023. Normality guided multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2665–2674.

Mandela Patrick, Yuki Asano, Polina Kuznetsova, Ruth Fong, Joao F Henriques, Geoffrey Zweig, and Andrea Vedaldi. 2020. Multi-modal self-supervision from generalized data transformations.

AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. 2020. Evolving losses for unsupervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 133–142.

Yujiang Pu and Xiaoyu Wu. 2022. Audio-guided attention network for weakly supervised violence detection. In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 219–223. IEEE.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. 2021. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. 1999. Support vector method for novelty detection. *Advances in neural information processing systems*, 12.

Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968.

Mahsa Shafaei, Christos Smailis, Ioannis A Kakadiaris, and Thamar Solorio. 2021. A case study of deep learning based multi-modal methods for predicting the age-suitability rating of movie trailers. *arXiv preprint arXiv:2101.11704*.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

V. C. Strasburger. 1989. Adolescent sexuality and the media. *Pediatric Clinics of North America*, 36(3):747–773.

Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. 2022. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

Vishaal Udandarao, Abhishek Maiti, Deepak Srivatsav, Suryatej Reddy Vyalla, Yifang Yin, and Rajiv Ratn Shah. 2020. Cobra: Contrastive bimodal representation algorithm. *arXiv preprint arXiv:2005.03687*.

Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. 2020. Self-supervised video representation learning by pace prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 504–521. Springer.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560.

Barbara J Wilson. 2008. Media and children's aggression, fear, and altruism. *The future of children*, pages 87–118.

Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. 2022a. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, pages 729–745. Springer.

Peng Wu and Jing Liu. 2021. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Transactions on Image Processing*, 30:3513–3527.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 322–339. Springer.

Peng Wu, Xiaotao Liu, and Jing Liu. 2022b. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia*.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. 2023. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. *arXiv preprint arXiv:2308.11681*.

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. 2019. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343.

Peng Xu, Xiatian Zhu, and David A Clifton. 2022. Multimodal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*.

Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. 2020. Video representation learning with visual tempo consistency. *arXiv preprint arXiv:2006.15489*.

Jianxi Yang, Xiaoxia Yang, Ren Li, Mengting Luo, Shixin Jiang, Yue Zhang, and Di Wang. 2023a. Bert and hierarchical cross attention-based question answering over bridge inspection knowledge graph. *Expert Systems with Applications*, page 120896.

Zekun Yang, Yuta Nakashima, and Haruo Takemura. 2023b. Multi-modal humor segment prediction in video. *Multimedia Systems*, pages 1–10.

Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. 2022. Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6278–6287.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004.

Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. 2023a. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16271–16280.

Litian Zhang, Xiaoming Zhang, and Junshu Pan. 2022. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11676–11684.

Yigeng Zhang, Mahsa Shafaei, Fabio Gonzalez, and Thamar Solorio. 2021. From none to severe: Predicting severity in movie scripts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3951–3956, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yigeng Zhang, Mahsa Shafaei, Fabio Gonzalez, and Thamar Solorio. 2023b. Positive and risky message assessment for music products. *arXiv preprint arXiv:2309.10182*.

Chen Zheng, Quan Guo, and Parisa Kordjamshidi. 2020. Cross-modality relevance for reasoning on language and vision. *arXiv preprint arXiv:2005.06035*.

Hang Zhou, Junqing Yu, and Wei Yang. 2023. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. *arXiv preprint arXiv:2302.05160*.

Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. 2022. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *arXiv preprint arXiv:2211.11275*.

9. Appendix

The appendix provides detailed explanations about pretraining approaches (9.1) and pretraining datasets (9.2) and the experimental setup (9.3) for both pre-training and fine-tuning processes.

9.1. Pretraining Approaches

In this section, we explain the pretraining approaches that we considered to warm-start our multimodal model.

9.1.1. Video-Audio-Text Matching

In our model, we pretrain for the tasks of Video-Text Matching (VTM), Video-Audio Matching (VAM), and Audio-Text Matching (ATM). We utilize three separate Multi-Layer Perceptron (MLP) blocks, for predicting VTM, ATM, and VAM. For each task, the corresponding representations of the two modalities are combined and processed through the MLP block. For instance, to predict ATM, the text and audio representations are concatenated and passed through the MLP block. A softmax function is employed to generate an output of either 0 or 1, where 0 signifies a mismatch between the modalities and 1 indicates a match.

In the pretraining process, HICCAP reads a batch of video/audio/text triplets from the dataset at each iteration. At this stage, all modalities are matched together, resulting in a matching label of 1. However, in order to create mismatch triplets, we employed a method of randomly generating mismatches among the batch samples. This method takes the batch of video, audio, and text vector triplets as input. Then for each sample (triplet) of a batch, it randomly selects one of the three modalities, and with a probability of p replaces it with the same modality from another sample in the batch. In case of replacement, the matching label of the modality and the other modalities would change to zero. As a result, when the selected modality is replaced, the label for the matching between that modality and the other modalities is set to zero. For example, if the video modality is chosen and then replaced with another video sample, the labels for VTM and VAM for the current sample would be 0, but the label for ATM would remain 1. This process is repeated for all samples in the batch, resulting in a new batch of samples with updated labels. During the training process, this function is called each time, and its output is used as the input for the Video-Text Matching, Video-Audio Matching, and Audio-Text Matching fully connected layers. During the pretraining, we compute the loss for Video-Text Matching (VTM), Video-Audio Matching (VAM), and Audio-Text Matching (ATM) tasks and aggregate all losses together to find the final loss.

$$\mathcal{L} = \lambda_{VTM}\mathcal{L}_{VTM} + \lambda_{VAM}\mathcal{L}_{VAM} + \lambda_{ATM}\mathcal{L}_{ATM} \quad (2)$$

9.1.2. Multimodal Contrastive Learning

Inspired by (Akbari et al., 2021; Gurram et al.), we utilize common space projection in conjunction with contrastive learning to effectively train our model. Specifically, we leverage the video, audio, and text representation outputs generated from the hierarchical-cross-attention part to establish a semantically common space mapping. We represent the embedding for modality m as z_m. For this purpose, we define the projection function $g_{m,m'}$ to project both z_m and $z_{m'}$ into a multi-modal latent space m and m'. Within these multi-modal latent spaces, we utilize a contrastive framework to compare embeddings using a cosine similarity function. Consequently, for all i, when $m' \neq m$, we achieve high similarity between $g_{m,m'}(z_{m,i})$ and $g_{m,m'}(z_{m',i})$, and for all $i \neq j$ when $m' \neq m$, we obtain low similarity $g_{m,m'}(z_{m,i})$ and $g_{m,m'}(z_{m',j})$. We use a linear projection (Three linear layers with ReLU and batch normalization) for the $g_{m,m'}$ mapping.

Cross-Modal Contrastive Loss: We illustrate the positive training pairs as various modalities from the same sample, while negative training pairs are generated from different modalities of distinct samples within a batch. Subsequently, a minibatch containing N video samples resulted in N positive pairs and $N^2 - N$ negative pairs. Utilizing Noise Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2010) as our contrastive loss, we aim to enhance the similarity between positive pairs and increase the dissimilarity between negative pairs in the associated joint embedding space. The NCE loss is expressed as follows (omitting the projection functions for simplicity):

$$NCE(z_m, z'_m) = -\log\left(\frac{\Sigma_{i=0}^{N}\exp(z_{m,i}^T z_{m',i}/\tau)}{\Sigma_{i=0}^{N}\exp(z_{m,i}^T z_{m',i}/\tau) + \Sigma_{i=0}^{N}\Sigma_{j\neq i}^{N}\exp(z_{m,i}^T z_{m',j}/\tau)}\right) \quad (3)$$

In accordance with (Chen et al., 2021a), we apply NCE to audio-video, audio-text, and video-text pairs:

$$\mathcal{L}_{AV}(z_a, z_v) = NCE(g_{av}(z_a), g_{av}(z_v)) \quad (4)$$

$$\mathcal{L}_{AT}(z_a, z_t) = NCE(g_{at}(z_a), g_{at}(z_t)) \quad (5)$$

$$\mathcal{L}_{VT}(z_v, z_t) = NCE(g_{vt}(z_v), g_{vt}(z_t)) \quad (6)$$

By aggregating all losses, we establish the complete loss for this phase of pretraining:

$$\mathcal{L} = \lambda_{A,V}\mathcal{L}_{AV} + \lambda_{A,T}\mathcal{L}_{AT} + \lambda_{V,T}\mathcal{L}_{VT} \quad (7)$$

where $\lambda_{m,m'}$ corresponds to the weight for the modality pair m and m' that is a learnable parameter and will be tuned during training.

9.2. Pretraining Datasets

We pretrain our model on various datasets containing video clips with corresponding audio clips and descriptions. To increase the versatility of our model, we use a diverse set of domains, such as human actions, movies, and personal collections. For this purpose, we use **Kinetics-400** dataset (Kay et al., 2017) and part of **HowTo100M** dataset (Miech et al., 2019), each with a different distribution of video, audio, and description.

Kinetics-400 Dataset (Kay et al., 2017) includes 400 diverse human action classes, each featuring at least 400 video clips. These clips, which have a duration of around 10 seconds, are extracted from individual YouTube videos. The variety of action categories spans from human-object interactions, such as playing musical instruments, to human-human interactions like shaking hands.

HowTo100M Dataset (Miech et al., 2019) is a huge collection of YouTube videos focused on instructional videos where content producers teach complicated activities with the specific goal of demonstrating the visual information displayed on the screen. It has 136M video clips with captions which include 23K activities from various domains extracted from 1.2M YouTube videos. Each video has a narration accessible as automatically downloaded subtitles from YouTube.

9.3. Experimental Setup

In this study, we employed the AdamW optimizer with a weight decay of 0.02, betas set at the default values of (0.9,0.999), and eps of 1e-8. Additionally, we used Pytorch's adaptive scheduler, which adjusts the learning rate by a factor of 0.5 and maintains a minimum learning rate of 1e-8. Our experiments were conducted on an A100 GPU with 40GB of memory, allowing for a batch size of 16. We performed 30 epochs, saving the optimal model for further use. Utilizing the A100 GPU and a batch size of 16, each pretraining epoch took approximately 2960 seconds (2630 seconds for training and 330 seconds for validation). In the fine-tuning phase, each epoch required around 250 seconds (220 seconds for training and 30 seconds for validation). Our implementation utilized Python 3.11.1, scikit-learn 1.2, and torchvision 0.14.1, along with NVIDIA-SMI 515.48.07, Driver Version 515.48.07, and CUDA Version 11.7.