ABSTRACT

Title of Dissertation: DEVELOPING MULTIMODAL LEARNING METHODS FOR VIDEO UNDERSTANDING

Mingwei Sun,
Doctor of Philosophy, 2024

Dissertation Directed by: Professor Kunpeng Zhang,
Department of Decision, Operations and Information Technologies

In recent years, the field of deep learning, with a particular emphasis on multimodal representation learning, has experienced significant advancements. These advancements are largely attributable to groundbreaking progress in areas such as computer vision, voice recognition, natural language processing, and graph network learning. This progress has paved the way for a multitude of new applications. The domain of video, in particular, holds immense potential. Video is often considered the most potent form of digital content for communication and the dissemination of information. The ability to effectively and efficiently comprehend video content could prove instrumental in a variety of downstream applications. However, the task of understanding video content presents numerous challenges. These challenges stem from the inherently unstructured and complex nature of video, as well as its interactions with other forms of unstructured data, such as text and network data. These factors contribute to the difficulty of video analysis. The objective of this dissertation is to develop deep learning methodologies capable of understanding video across multiple dimensions. Furthermore, these methodologies aim to offer a degree of

interpretability, which could yield valuable insights for researchers and content creators. These insights could have significant managerial implications.

In the first study, I introduce an innovative network based on Long Short-Term Memory (LSTM), enhanced with a Transformer co-attention mechanism, designed for the prediction of apparent emotion in videos. Each video is segmented into clips of one-second duration, and pre-trained ResNet networks are employed to extract audio and visual features at the second level. I construct a co-attention Transformer to effectively capture the interactions between the audio and visual features that have been extracted. An LSTM network is then utilized to learn the spatiotemporal information inherent in the video. The proposed model, termed the Sec2Sec Co-attention Transformer, outperforms several state-of-the-art methods in predicting apparent emotion on a widely recognized dataset: LIRIS-ACCEDE. In addition, I conduct an extensive data analysis to explore the relationships between various dimensions of visual and audio components and their influence on video predictions. A notable feature of the proposed model is its interpretability, which enables us to study the contributions of different time points to the overall prediction. This interpretability provides valuable insights into the functioning of the model and its predictions.

In the second study, I introduce a novel neural network, the Multimodal Co-attention Transformer, designed for the prediction of personality based on video data. The proposed methodology concurrently models audio, visual, and text representations, along with their intra-relationships, to achieve precise and efficient predictions. The effectiveness of the proposed approach is demonstrated through comprehensive experiments conducted on a real-world dataset, namely, First Impressions. The results indicate that the proposed model surpasses state-of-the-art methods in performance while preserving high computational efficiency. In addition to evaluating the performance of the proposed model, I also undertake a thorough interpretability analysis to examine the contribution

across different levels. The insights gained from the findings offer a valuable understanding of personality predictions. Furthermore, I illustrate the practicality of video-based personality detection in predicting outcomes of MBA admissions, serving as a decision support system. This highlights the potential importance of the proposed approach for both researchers and practitioners in the field.

In the third study, I present a novel generalized multimodal learning model, termed VAN, which excels in learning a unified representation of **v**isual, **a**coustic, and **n**etwork cues. Initially, I utilize state-of-the-art encoders to model each modality. To augment the efficiency of the training process, I adopt a pre-training strategy specifically designed to extract information from the music network. Subsequently, I propose a generalized Co-attention Transformer network. This network is engineered to amalgamate the three distinct types of information and to learn the intra-relationships that exist among the three modalities, a critical facet of multimodal learning. To assess the effectiveness of the proposed model, I collect a real-world dataset from TikTok, comprising over 88,000 videos. Extensive experiments demonstrate that the proposed model surpasses existing state-of-the-art models in predicting video popularity. Moreover, I have conducted a series of ablation studies to attain a deeper comprehension of the behavior of the proposed model. I also perform an interpretability analysis to study the contributions of each modality to the model performance, leveraging the unique property of the proposed co-attention structure. This research contributes to the field by proffering a more comprehensive approach to predicting video popularity on short-form video platforms.

DEVELOPING MULTIMODAL LEARNING METHODS FOR VIDEO UNDERSTANDING

by

Mingwei Sun

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2024

Advisory Committee:
Professor Kunpeng Zhang, Chair
Professor P.K. Kannan
Professor Lauren Rhue
Professor Jessica Clark
Professor Vanessa Frias-Martinez, Dean's Representative

Acknowledgments

During the course of my six-year doctoral study at Smith Business School, I have had the privilege of collaborating with numerous esteemed researchers. Their guidance, teachings, and inspiration have been instrumental in my journey. I am deeply appreciative of their support and express my gratitude wholeheartedly.

First and foremost, I would like to extend my deepest gratitude to my advisor, Professor Kunpeng Zhang, for his unwavering support and valuable advice throughout my six years of study. His consistent provision of constructive feedback and advice has not only enriched my knowledge but also fostered my growth as an independent researcher. My academic journey under the guidance of Professor Zhang has been both enriching and transformative. I am deeply grateful for everything he has provided me with. His mentorship has truly been a cornerstone of my academic development.

I am sincerely thankful to Professor Jessica Clark, with whom I initiated my first project. The discussions about research, career, and life with her have been enlightening. I am also grateful to Professor Lauren Rhue for her consistent support and patience. Working with her has been a rewarding experience, and I have learned much from her. I also extend my gratitude to Professor P.K. Kannan for his insightful advice on both my research and career. I am thankful to Professor Vanessa Frias-Martinez for her unique perspective on research, which has significantly influenced my thinking process. My heartfelt thanks also go to Professor Balaji Padmanabhan,

whose guidance has been extremely beneficial. His insights on positioning my work have been enlightening. In general, I have learned a lot from these top-notch researchers.

Furthermore, I would like to acknowledge Justina and Miloyka for their immense help, encouragement, and patient support. They have always been available to assist, regardless of the challenges I faced. My doctoral experience has been productive, also due to the interactions with my peers, particularly Bingze Xu, Wei Feng, Gujie Li, Sung Hyun Kwon, Maya Mudambi, Feiyu E, Weihong Zhao, Yunfei Wang, among others.

Finally, I would like to express my deepest appreciation to my wife, Fan Yu, and my parents for their unconditional support, both mentally and financially. Their backing has been a pillar of strength in my journey.

Table of Contents

Acknowledgements ii

Table of Contents iv

List of Tables vi

List of Figures vii

Chapter 1: Introduction 1

Chapter 2: Sec2Sec Co-attention Transformer for Video Emotion Prediction 7
- 2.1 Introduction . . . 7
- 2.2 Related Work . . . 11
  - 2.2.1 Emotion and Its Impact . . . 12
  - 2.2.2 Audio-Video Representation Learning . . . 13
  - 2.2.3 Transformer and Its Application . . . 14
- 2.3 Preliminaries . . . 15
- 2.4 Method: Sec2Sec Co-attention Transformer . . . 17
- 2.5 Experiments . . . 21
  - 2.5.1 Dataset . . . 21
  - 2.5.2 Implementation Details . . . 23
  - 2.5.3 Baselines . . . 23
  - 2.5.4 Results . . . 26
- 2.6 Ablation Study . . . 27
- 2.7 Conclusion . . . 28

Chapter 3: Multimodal Co-attention Transformers for Video-Based Apparent Personality Understanding 30
- 3.1 Introduction . . . 30
- 3.2 Related Work . . . 36
  - 3.2.1 Personality . . . 36
  - 3.2.2 Video-Based Deep Personality Prediction . . . 38
  - 3.2.3 Transformers in Computer Vision . . . 40
  - 3.2.4 Multimodal Learning . . . 41
- 3.3 Preliminaries . . . 42
- 3.4 Our Model . . . 43
- 3.5 Experiments . . . 49

3.5.1 Dataset . . . 49
3.5.2 Implementation Details . . . 50
3.5.3 Baselines . . . 51
3.5.4 Evaluation Metrics . . . 52
3.5.5 Efficiency Metric . . . 53
3.6 Results . . . 53
3.6.1 Performance . . . 53
3.6.2 Efficiency . . . 55
3.7 Ablation Study . . . 58
3.8 Interpretability Analysis . . . 58
3.9 Decision Support Showcasing: MBA Admission Prediction . . . 62
3.10 Discussion and Conclusion . . . 68

Chapter 4: Network-enhanced Multimodal Co-attention Learning for Short-Form Video Popularity Prediction 72
4.1 Introduction . . . 72
4.2 Related Work . . . 76
4.2.1 Online Video Popularity Prediction . . . 76
4.2.2 Multimodal Representation Learning . . . 77
4.3 Proposed Model . . . 79
4.3.1 Input Embeddings . . . 80
4.3.2 VAN: A Generalized Multimodal Co-attention Network . . . 83
4.4 Experiments . . . 86
4.4.1 Dataset . . . 86
4.4.2 Popularity Score . . . 86
4.4.3 Evaluation Metrics . . . 87
4.4.4 Implementation Details . . . 88
4.4.5 Graph Attention Network Pre-training . . . 89
4.4.6 Baselines . . . 90
4.5 Results . . . 91
4.6 Ablation Study . . . 93
4.7 Interpretability Analysis . . . 96
4.8 Conclusion . . . 99

Appendix A: Video Essay Recording Page 101

Appendix B: Additional Interpretability Analysis 102

Bibliography 104

List of Tables

2.1 The t-test comparison of audio features on LIRIS-ACCEDE between high emotional and low emotional videos. . . . 9
2.2 The t-test comparison of visual features on LIRIS-ACCEDE between high emotional and low emotional videos. . . . 10
2.3 Performance comparison of our model with baselines for arousal prediction. . . . 23
2.4 Performance comparison of our model with baselines for valence prediction. . . . 24
2.5 Comparison of batch vs. layer normalization. . . . 28

3.1 Summary Statistics for First Impressions . . . 50
3.2 Performance comparison of our model with baselines for Big-Five personality predictions. . . . 55
3.3 Cost Analysis . . . 56
3.4 Comparison of Positional Encoding vs No Positional Encoding . . . 58
3.5 Summary Statistics for the Case Study of MBA Admission . . . 63
3.6 Personality Trait Correlations . . . 63
3.7 Estimation Results for MBA Admission . . . 64
3.8 Factor Analysis . . . 66
3.9 Estimation Results for MBA Admission Using Extracted Factor . . . 67

4.1 Hyperparameters of the proposed model. . . . 88
4.2 Performance comparison of our model with baselines. . . . 91
4.3 Ablation study of the proposed model. . . . 93

List of Figures

2.1 An illustration example of images eliciting different emotions. 8
2.2 An illustration example of various emotions elicited by different audio waveforms. 9
2.3 Correlation heatmaps between audio and visual features for two emotional states and two emotional intensities. 11
2.4 An overview of our proposed model: Sec2Sec Co-attention Transformer. 17
2.5 The co-attention block. 19
2.6 The Sec2Sec Structure. 22
2.7 The LSTM Attention. 27
2.8 Effect of key hyperparameters of Sec2Sec SA-CA on accuracy and $F1$ score. . . 27

3.1 Images sampled from videos in the First Impressions dataset that exhibit varying degrees of personality traits. 31
3.2 Audio waveforms in the First Impressions dataset that exhibit varying degrees of personality traits. 32
3.3 An overview of our proposed model: Multimodal Co-attention Transformer. . . . 44
3.4 Average modality importance on personality prediction. 59
3.5 Average contributions of images of different positions on personality prediction. . 60
3.6 Region contributions of images on personality predictions. 60
3.7 Parallel Analysis Scree Plots. 66

4.1 An illustration of a TikTok post. 73
4.2 An overview of our proposed model: A Generalized Multimodal Co-attention Transformer. 79
4.3 Spearman's Rank Correlations regarding the number of sampled frames. 96
4.4 The average contributions of each modality. 98

A.1 A screenshot of the recording page for the video essay. 101

B.1 The contributions of each modality for Head 1. 102
B.2 The contributions of each modality for Head 2. 103
B.3 The contributions of each modality for Head 3. 103
B.4 The contributions of each modality for Head 4. 103

Chapter 1: Introduction

In the past decade, video content has emerged as an integral component of people's daily routines. As of 2023, individuals are allocating an average of 17 hours per week to the consumption of online video content [78]. Specifically, in the United States, it is anticipated that by 2024, there will be 164.6 million internet users engaging with video content, as per data from Statistica [18]. This surge in video content consumption has led to a transformative impact on social media platforms. A majority of these platforms, including TikTok, Instagram, and Facebook, now offer video-sharing services. To illustrate, TikTok reported 1.04 billion monthly active users in May 2024 [92], with an annual expenditure of $3.84 billion from consumers [19]. Given this context, video marketing presents immense business potential. In fact, 91% of businesses are leveraging video as a marketing tool [106].

The ubiquity and exponential growth of video marketing have sparked a surge of interest among scholars and industry professionals alike. They are keen to comprehend the multifaceted dimensions of video content, including auditory intensity (loudness) [40], aesthetic considerations (color choice) [56], and verbal (topic) and non-verbal (facial expression) communication cues [55]. While these elements provide valuable insights, they are relatively straightforward to extract and interpret. They represent only the surface-level understanding of video content, leaving a vast array of deeper, more complex aspects unexplored. These uncharted territories hold the potential

to unlock a more profound understanding of video content, thereby paving the way for a myriad of downstream analyses and studies.

In the sphere of artificial intelligence, the past few years have witnessed remarkable strides in the field of deep learning, with a particular emphasis on multimodal learning. This progress has been fueled by groundbreaking advancements in several sub-domains, including computer vision, natural language processing, voice recognition, and network learning. As a result, multimodal learning has demonstrated exceptional performance across a wide array of applications. One such application is video analytics, where videos typically comprise visual and auditory components. This makes video analytics a natural and fitting application for multimodal learning.

Despite the immense potential of multimodal learning in this domain, it is not without its challenges. The application of existing methods may not always yield optimal performance in a given context due to several reasons. Firstly, capturing the interaction and alignment between the auditory and visual components of a video is a complex task. I posit that this aspect is integral to the performance of the model. Secondly, temporal information embedded within the video is of critical importance. The ability to accurately utilize this information can significantly enhance the model's performance. Thirdly, the fusion of auditory and visual data with other types of information, such as text and graph networks, is a crucial consideration. The integration of these diverse data types can provide a more holistic understanding of the video content. Lastly, the current practice of employing deep learning models in video understanding is predominantly a black-box approach. This lack of interpretability impedes the broader adoption of these powerful models and fails to provide actionable insights and managerial implications.

In light of this, the primary objective of this dissertation is to delve deeper into the realm of video understanding. To achieve this, I propose the development and application of advanced

deep learning methodologies. These methods are designed to penetrate beyond the superficial layers of video content, enabling a more comprehensive and nuanced understanding. By harnessing the power of multimodal learning techniques, I can uncover hidden patterns and alignments within the video content. This, in turn, can facilitate a wide range of applications, from enhancing the effectiveness of video marketing strategies to improving the accuracy of video-based predictive models. Ultimately, this research aims to contribute significantly to the field of video content analysis, setting new benchmarks for future studies in this domain.

Study 1: Sec2Sec Co-attention Transformer for Video Emotion Prediction

Video-based apparent emotion detection plays a crucial role in video understanding, as they encompass various elements such as vision, audio, audio-visual interactions, and spatiotemporal information, which are essential for accurate video predictions. However, existing approaches often focus on extracting only a subset of these elements, resulting in the limited predictive capacity of their models. To address this limitation, I propose a novel LSTM-based network augmented with a Transformer co-attention mechanism for predicting apparent emotion in videos. Specifically, I divide each video into one-second clips and utilize pre-trained ResNet networks to extract audio and visual features at the clip level. I develop a co-attention Transformer to effectively capture the interactions between the extracted audio and visual features and leverage an LSTM network to learn the spatiotemporal information present in the video. I demonstrate that the proposed Sec2Sec Co-attention Transformer surpasses multiple state-of-the-art methods in predicting apparent emotion on a widely used dataset: LIRIS-ACCEDE. Additionally, I perform comprehensive data analysis to investigate the relationships between different dimensions of visual and audio components and their impact on video predictions. Notably, my model offers interpretability, allowing me to examine the contributions of different time points to the overall

prediction.

Study 2: Multimodal Co-attention Transformers for Video-Based Apparent Personality Understanding

Video has emerged as a pervasive medium for communication, entertainment, and information sharing. With the consumption of video content continuing to increase rapidly, understanding the impact of visual narratives on personality has become a crucial area of research. While text-based personality understanding has been extensively studied in the literature, video-based personality prediction remains relatively under-explored. Existing approaches to video-based personality prediction can be broadly categorized into two directions: learning a joint representation of audio and visual information using fully-connected feed-forward networks, and separating a video into its individual modalities (text, image, and audio), training each modality independently, and then ensembling the results for subsequent personality prediction. However, both approaches have notable limitations: ignoring complex interactions between visual and audio components, or considering all three modalities but not in a joint manner. Furthermore, all methods require high computational costs as they require high-resolution images to train. In this chapter, I propose a novel Multimodal Co-attention Transformer neural network for video-based personality prediction. My approach simultaneously models audio, visual, and text representations, as well as their inter-relations, to achieve accurate and efficient predictions. I demonstrate the effectiveness of my method via extensive experiments on a real-world dataset: First Impressions. My results show that the proposed model outperforms state-of-the-art approaches while maintaining high computational efficiency. In addition to my performance evaluation, I also perform a set of comprehensive interpretability analyses to investigate the contribution across different levels. My findings reveal valuable insights into personality predictions. In addition, I showcase the utility of

video-based personality detection in predicting MBA admission outcomes as a decision support system, highlighting its potential significance for both researchers and practitioners.

Study 3: Network-enhanced Multimodal Co-attention Learning for Short Video Popularity Prediction

The recent surge in the popularity of short-form videos has unveiled considerable opportunities for business applications, encompassing personalized recommendations and targeted advertising. Predominantly, traditional research employs acoustic-visual information for making predictions about video popularity. However, a unique feature of these platforms is the music network, which provides an abundance of information on the distribution and sharing of various trending songs. This could potentially influence video popularity, a factor often overlooked in existing literature. In this chapter, I introduce a novel generalized multimodal learning model, termed VAN, which is adept at learning a unified representation of **v**isual, **a**coustic and **n**etwork cues. Initially, we employ cutting-edge encoders to model each modality. To enhance the efficiency of the training process, I design a pre-training strategy specifically tailored to extract information from the music network. As a final step, I put forward a generalized Co-attention Transformer network. This network is designed to fuse the three distinct types of information and to learn the intra-relationships that exist among the three modalities, a crucial aspect of multimodal learning. To evaluate the effectiveness of the proposed model, I have collected a real-world dataset from TikTok, consisting of over 88,000 videos. My comprehensive experiments demonstrate that my model outperforms existing state-of-the-art models in predicting video popularity. Furthermore, we have conducted a series of ablation studies to gain a deeper understanding of the behavior of my model. We additionally conduct interpretability analysis to examine the contributions of each modality to the model performance by leveraging the distinctive property of the proposed

co-attention structure. This research contributes to the field by offering a more comprehensive approach to predicting video popularity on short-form video platforms.

Collectively, the proposed methods and the findings of three studies in this dissertation provide us with a more profound comprehension of videos from multiple perspectives. This enhanced understanding paves the way for a plethora of downstream research opportunities, thereby expanding the horizons of knowledge in video analytics. The insights gained from these studies provide valuable guidance for both researchers and practitioners. For researchers, these insights can inform the design of future studies, helping to refine research questions, hypotheses, and methodologies. For practitioners, particularly those in the realm of video marketing and analytics, these insights can inform strategic decision-making, helping to optimize the effectiveness of video content and drive business outcomes.

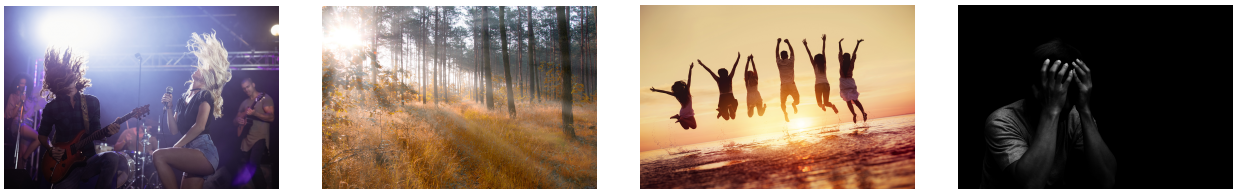
Chapter 2: Sec2Sec Co-attention Transformer for Video Emotion Prediction

2.1 Introduction

Emotions are generally described as mental states by neurophysiological changes. They can be associated with thoughts, feelings, behavioral responses, and a degree of pleasure or displeasure [104], which can accordingly affect our decision-making and eventually shape how we perceive the world ubiquitously. As cognitive processes can be profoundly influenced by emotions [77], people primarily rely on emotional levels when making their judgments [11]. Several dimensions that could be linked to emotion-related responses have been identified [73]. Among these two major ones have been widely explored in the literature. They are pleasure-displeasure and arousal-sleep dimensions. Specifically, The former indicates the degree of positivity or negativity of the experience, also known as valence, while the latter assesses the level of energy or fatigue that an experience produces, also known as arousal.

There has been a growing interest in understanding human emotions from both researchers and practitioners. Emotion detection has been the main focus of existing literature, which is also the scope of our study. Various methods have been proposed to detect emotions, especially for text documents. For example, Kratzwald et al. developed a transfer learning-based model (called *sent2affect*) for emotion recognition [46]. Su et al. designed a long short-term memory (LSTM) network to predict emotions based on the combination of semantic and emotional words

Figure 2.1: An illustration example of images eliciting different emotions.



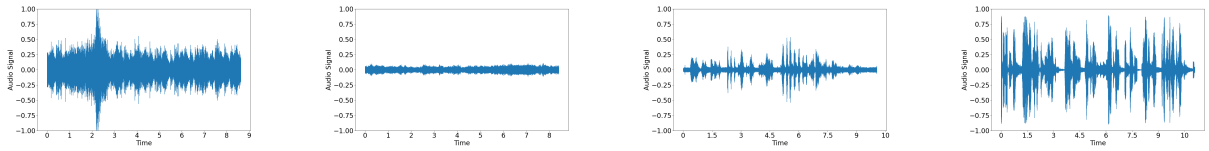
(a) High Arousal (b) Low Arousal (c) High Valence (d) Low Valence

Note: A high-arousal emotion is evoked by a violet background and an exciting concert in Figure 2.1a. A combination of light colors and a calm and peaceful natural environment in Figure 2.1b conveys a low-arousal emotion. A warm background along with happy individuals in Figure 2.1c induce a high-valence emotion. A dark background and a sad person in Figure 2.1d provoke a low-valence emotion.

[81]. However, video-based emotion detection remains under-explored, even though videos have been largely generated and posted on various platforms. Prior studies found that videos are more efficient and effective to elicit emotions compared to text [46].

A video often consists of two components: vision and audio. Both can stimulate emotions in their own ways. Figure 2.1 shows different images can elicit different emotions, as suggested by [43]. On the other hand, [103] has shown that the design of electronic music and audio-visual media can elicit audience emotions. In particular, non-diegetic music is a key feature that elicits states of emotions. Audio is usually represented by a continuous waveform. Audio waveform measures how sound pressure varies over time. Figure 2.2 shows different patterns of audio waveforms that can convey different states of emotions. Furthermore, we examine a set of audio and visual features that exhibit significant differences between videos with high and low emotional intensities by performing t-test analyses on the LIRIS-ACCEDE dataset [10]. Table 2.1 and Table 2.2 are the t-test comparisons of audio and visual features across videos with high and low emotional intensities. From the tables, we can see that many visual and audio features do play a role in distinguishing video emotions.

Figure 2.2: An illustration example of various emotions elicited by different audio waveforms.



(a) High Arousal (b) Low Arousal (c) High Valence (d) Low Valence

Note: A high-arousal emotion is provoked by a spike with high sound pressure in Figure 2.3a. A low-arousal emotion is conveyed by low sound pressure over time in Figure 2.3b. Few spikes with small sound pressure in Figure 2.3c elicit a high-valence emotion. A low-valence emotion is evoked by multiple spikes with high sound pressure in Figure 2.3d.

Table 2.1: The t-test comparison of audio features on LIRIS-ACCEDE between high emotional and low emotional videos.

Feature	Emotion	p-value (significant?)
Pitch	valence	1.84e-15(✓)
Pitch	arousal	4.31e-05(✓)
AL	valence	0.61(✗)
AL	arousal	5.74e-53(✓)
Spectral rolloff	valence	3.24e-14 (✓)
Spectral rolloff	arousal	0.01(✓)
ZRC	valence	7.36e-09(✓)
ZRC	arousal	0.024(✓)

Note: Pitch means the average pitch of an audio sample. AL means the average loudness. Spectral roll-off means the frequency that holds a certain percentage of energy (e.g. 85%). Zero-crossing rate (ZRC) means the frequency that a waveform crosses zero representing audio smoothness.

Vision and audio can also jointly affect emotions. Specifically, if a visual component is well-aligned with its counterpart audio in the same time frame (e.g., within a second), it can create a synergy that stimulates emotions more intensely. A good alignment could be a high degree of consistency or correspondence between two components. Most importantly, eliciting different emotional states requires different combinations of audio and visual features. For example, a cold picture with a low audio tempo is likely to stimulate a negative valence. Low arousal is conveyed

Table 2.2: The t-test comparison of visual features on LIRIS-ACCEDE between high emotional and low emotional videos.

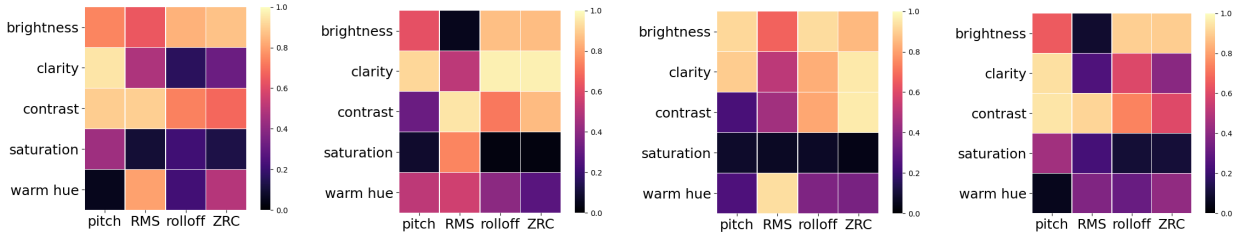
Feature	Emotion	p-value (significant?)
saturation	valence	1.40e-19 (✓)
saturation	arousal	1.04e-05 (✓)
brightness	valence	1.38e-39 (✓)
brightness	arousal	4.36e-10 (✓)
contrast	valence	4.44e-13 (✓)
contrast	arousal	0.76 (✗)
clarity	valence	9.38e-17 (✓)
clarity	arousal	0.005 (✓)
warm hue	valence	2.11e-55 (✓)
warm hue	arousal	4.17e-16 (✓)

Note: Saturation indicates average saturation across pixels. Brightness: average intensity across pixels. Contrast indicates the standard deviation of intensity across pixels. Clarity indicates the proportion of pixels with intensity above a certain threshold (e.g. 0.7). Warm hue indicates the proportion of warm colors in a frame.

by a piece of light music and a peaceful environment. For the sake of illustration, the correlation heatmap[1] between audio and visual features for various emotional states and intensities are plotted in Figure 2.3 using LIRIS-ACCEDE [10], which confirms that the integration of audio and visual signals induces various emotional states and intensities. Another important factor that might affect how people perceive a video emotionally is the sequential composition of a video. The temporal pattern exhibited in a video should be captured for emotion prediction. For example, people are more likely to recall the most recently presented information [61], which indicates that later audio clips may be weighted higher when estimating emotions.

Despite the popularity of videos and the practical importance of detecting their emotions, very limited research has been conducted to quantitatively estimate how videos induce emotions. In this paper, we are among the pioneers to fill this research gap by developing a Transformer-

[1]The correlations are computed by performing the following steps: (1) splitting each video into n one-second clips; (2) extracting audio and visual features from each one-second clip; (3) calculating the correlations between those features for each video; and (4) averaging each correlation pair.



(a) High Arousal (b) Low Arousal (c) High Valence (d) Low Valence

Figure 2.3: Correlation heatmaps between audio and visual features for two emotional states and two emotional intensities.

based second-to-second (Sec2Sec) co-attention model to predict the perceived emotional states of videos when we watch them. Specifically, we first implement a Transformer-based co-attention network extended from the work proposed by [22] to understand the interaction between audio and visual components. We further combine an LSTM module with such a co-attention network to capture the temporal information of videos at the second level. To do so, we first split each video into one-second video clips. We then feed each one-second clip into our designed co-attention network. The output of each video clip from the co-attention network is fed into an LSTM network sequentially. Lastly, we add a fully-connected network to predict emotions. To evaluate our work, we conduct experiments on a real-world dataset from LIRIS-ACCEDE with $9,800$ videos [10]. The experimental results show that our model outperforms several cutting-edge baselines, in terms of $F1$-score for both arousal and valence.

2.2 Related Work

Our work is closely related to three streams of literature: emotion, audio-video representation learning and applications of Transformers.

2.2.1 Emotion and Its Impact

Emotion, a multifaceted psychological phenomenon, lacks a universally acknowledged definition. However, it is frequently characterized as a mental condition that incorporates cognitive processes, affective states, physiological alterations, and behavioral reactions, all of which are marked by varying intensities of pleasure or displeasure [65]. The genesis of emotions can be attributed to a multitude of mechanisms. These include bottom-up processes initiated by external stimuli, top-down processes that involve the cognitive evaluation and interpretation of events based on accumulated experience and knowledge, or an amalgamation of both [59].

Numerous models have been put forth to delineate the dimensions of emotion. Ekman's theory of basic emotions [32] is one such model, which advocates the existence of six universal emotions. Another model, proposed by Cordaro et al. [25], extends Ekman's basic emotion theory to identify 22 distinct emotions. These prevalent models primarily concentrate on discerning the emotions experienced by the individual. However, in the realm of entertainment, such as social media, the influencer's focus shifts towards the emotions experienced by the audience. In this paper, we employ the widely recognized circumplex model of emotion or affect [74] to scrutinize the emotions elicited in the audience in response to stimuli. The circumplex model characterizes emotions along two dimensions: valence and arousal. Valence signifies the sentiment associated with an experience, spanning from pleasant to unpleasant. Arousal denotes the degree of intensity or activation associated with the experience, ranging from low to high. Valence and arousal encapsulate the pleasantness and intensity of the video experiences from the audience's viewpoint.

A considerable volume of research exists that examines the phenomenon of consumer

emotion and its consequential impact on consumer behavior. It is well-documented that consumers are susceptible to the influence of others' emotional expressions [38]. For instance, research has demonstrated that positive facial expressions in fundraising advertisements can sway funding decisions in a beneficial direction [69]. Furthermore, emotions encapsulated in online product reviews can markedly affect the perceived utility of the information [110]. Elements of the broader context, such as culture, can also mold how consumers react to others' emotions. Consumers of European cultural descent respond more robustly to excited expressions, while consumers of Chinese descent exhibit a stronger response to calm emotional expressions [66]. A majority of these studies explore emotions in text-based or visual mediums such as online reviews, social media, or facial expressions [13, 69, 80]. However, despite the richness of video content, there is limited research investigating the impact of perceived emotion, primarily due to the absence of a predictive model capable of forecasting perceived video emotion. Consequently, our study endeavors to bridge this gap.

2.2.2 Audio-Video Representation Learning

Audio Representation Learning. Traditionally, research often adopts hand-crafted audio feature extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs) [27]. Extracting MFCCs is an audio processing technique that models how human ears sense and resolve sound frequencies [71]. Recently, with the development of deep learning, researchers have explored several audio autoencoder techniques. Tagliasacchi et al. applied a convolutional deep belief network on music and speech data to solve different classification tasks [89]. Cartwright et al. designed a network, including an audio sub-network and a temporal network, to predict the long-

term and cyclic temporal structure using self-supervision [17]. Chung et al. explored a sequence-to-sequence autoencoder by incorporating RNN and LSTM together [23].

Audio-Visual Cross-Modal Learning. As videos provide a natural bridge between audio and visual events, they tend to happen together. Therefore, the mainstream of audio-visual representation learning research is to predict the synchronization or correspondence of audio and visual streams in videos. Arandjelovic and Zisserman trained an audio-visual cross-modal network from scratch to predict video correspondence [8]. Alwassel et al. used one clustered modality as a supervisory signal for another modality and predicted correspondence between two modalities [7]. Cheng et al. further developed three self-supervised co-attention-based networks to discriminate visual events related to audio events [22]. In addition, Kuhnke et al. proposed a two-stream aural-visual model (AVM) to predict facial expressions in videos [47].

2.2.3 Transformer and Its Application

Transformers in Natural Language Processing. Transformer was first introduced in the task of machine translation [97]. It has been a state-of-the-art natural language processing (NLP) architecture ever since. A variety of Transformer-based models have been developed to address many NLP tasks, mainly focusing on two streams. One follows the trend of pre-training Transformer-based models on large corpora and fine-tuning parameters on downstream NLP tasks. BERT is a pioneer that employs a multi-layer bi-directional Transformer model architecture [29]. However, BERT-based models are only capable of handling 512 tokens, which is not enough for long text documents. Hence, Longformer extends BERT while utilizing sliding window, dilated sliding window, and global attention to handle long text documents [12]. Unlike BERT-based models,

another stream of research focuses on language modeling as a pre-training task, such as GPT [70]. GPT was developed for text generation tasks such as question answering, text summarization, and many others, which has achieved great performance on downstream tasks in zero-shot or few-shot settings.

Transformers in Computer Vision. Recently, there has been increasing attention on applying Transformers to computer vision (CV) tasks as an alternative to convolutional neural networks (CNN). Many studies have achieved great success. ViT applies a Transformer model to linearly projected sequences of image patches to classify full images [31]. Swin Transformer improves ViT by introducing a hierarchical Transformer architecture and a shifted window scheme [53]. These are two representative Transformer-based models for image classification tasks. In order to classify video tasks, ViViT extends ViT by proposing two methods for embedding video samples: uniform frames sampling and tubelet embedding, and four model variants based on Transformer: spatiotemporal attention, factorized encoder-decoder, factorized self-attention, and factorized dot product attention [9]. Video Swin Transformer further extends Swin Transformer by introducing a 3D-shifted window-based multi-head self-attention module and a locality inductive bias to the self-attention module [54]. All these video-based analyses do not separate vision and audio and explicitly learn the joint effect on subsequent tasks, which is our focus in this study.

2.3 Preliminaries

In this section, we first briefly explain how multi-head self-attention works. It is the key component in Transformer that maps a query vector Q, a key vector K and a value vector V to an output

(embedding), as shown in Equation 3.1.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{2.1}$$

Specifically, self-attention is achieved by computing the dot product of Q and K divided by the square root of the dimension of Q denoted by d, which gives the similarity scores between Q and K, which is also known as the scaled-dot product. The scores are then translated into probabilities by applying a *softmax* function. Lastly, the probabilities are multiplied by V to get the final output for the next layer. The basic idea of the self-attention mechanism is to focus more on the vectors with high probabilities in V in the following layers. However, one self-attention layer limits the model's ability to focus on more positions without compromising other positions. To mitigate this limitation, it introduces a multi-head attention mechanism, which can increase the overall model performance. Specifically, the multi-head attention layer consists of h paralleled self-attention sub-layers, called "heads". Each head learns different query, key and value matrices. Different heads project input features into different sub-spaces. The output features from each head are concatenated into one matrix for the following layers, as shown in Equation 3.2.

$$MultiHead(Q, K, V) = Concat(h_1, h_2, ..., h_h)W^O$$
$$where\ h_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.2}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_{model}}$ are learnable weights for each head i. $W^O \in \mathbb{R}^{h \times d_{model} \times d_{model}}$ is the projection weight.

A fully-connected feed-forward network is applied after the self-attention layer. In addition,

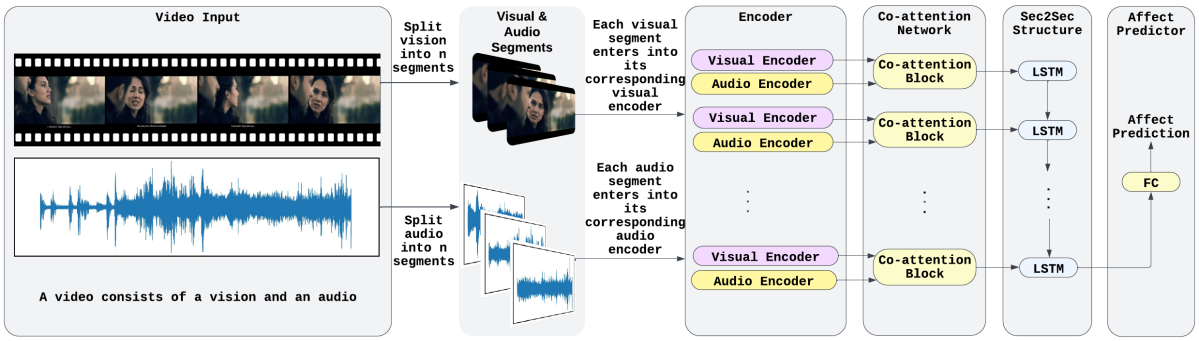


Figure 2.4: An overview of our proposed model: Sec2Sec Co-attention Transformer.

a residual connection and a normalization function are applied to each sub-layer.

2.4 Method: Sec2Sec Co-attention Transformer

This section presents our proposed model, which considers visual and audio representations, their interactions and the temporal information of videos. As depicted in Figure 4.2, the proposed model consists of five components: **Video segmentation**: We first split each video into n video segments. For instance, each segment consists of a one-second visual component and a one-second audio component. **Encoder network**: The encoder network comprises a visual encoder and an audio encoder that extract visual and audio features using pre-trained ResNet networks. [39]. **Co-attention block**: The co-attention block leverages Transformer [97] to model the interactions between visual and audio features, shown in Figure 2.5. **Sec2Sec structure**: It captures the temporal information via an LSTM network, illustrated in Figure 2.6. **Predictor**: The output from the LSTM network is fed into a fully-connected feed-forward network to make emotion predictions.

Visual encoder. To extract visual features, we first sample m frames per segment. Each frame is represented by a color image with the Red-Green-Blue (RGB) channels. Like prior studies,

we apply pre-processing to images, such as resizing to a dimension of 80x80, center cropping to 64x64, and normalizing based on the mean of (0.485, 0.456, 0.406) and the standard deviation of (0.229, 0.224, 0.225). Thus, each visual part is represented in a 4-dimensional space (i.e., 3-dimensional RGB plus m frames), which is fed into a pre-trained R(2+1)D ResNet model. R(2+1)D ResNet [93] is an extension of ResNet, utilizing 3D convolution and 3D pooling to learn the temporal features of videos.

Audio Encoder. For the audio segment, we first compute Mel-Frequency Cepstral Coefficients (MFCCs) [27], MFCC's first-order (delta coefficients), and second-order frame-to-frame time derivatives (delta-delta coefficients) from each audio clip. MFCCs are the coefficients that model how human ears sense and resolve sound frequencies [71]. The delta coefficients are used to capture speech rate information. The delta-delta coefficients are used to measure the acceleration of speech. Both delta coefficients and delta-delta coefficients are jointly used to measure the temporal information of an audio signal [71]. Each of the three coefficients is 2-dimensional. Therefore, the audio feature can be represented by combining three-channel MFCC features in which each channel is one type of coefficient. The extracted three-channel MFCC features are fed into a pre-trained ResNet [39]. ResNet introduces an identity shortcut connection to solve the vanishing gradient problem, which outperforms other CNN models on popular image classification tasks. In our work, the 3-channel MFCC audio features are considered a special type of "image". Hence, we use a pre-trained 18-layer ResNet to obtain the audio features.

Co-attention block. As illustrated in Figure 2.5, the extracted visual and audio features for each segment enter into two symmetrical co-attention sub-blocks, visual and audio sub-blocks, to learn guided audio and visual representations. Each sub-block is built by combining a standard multi-head self-attention module with a multi-head co-attention module. A normalization layer

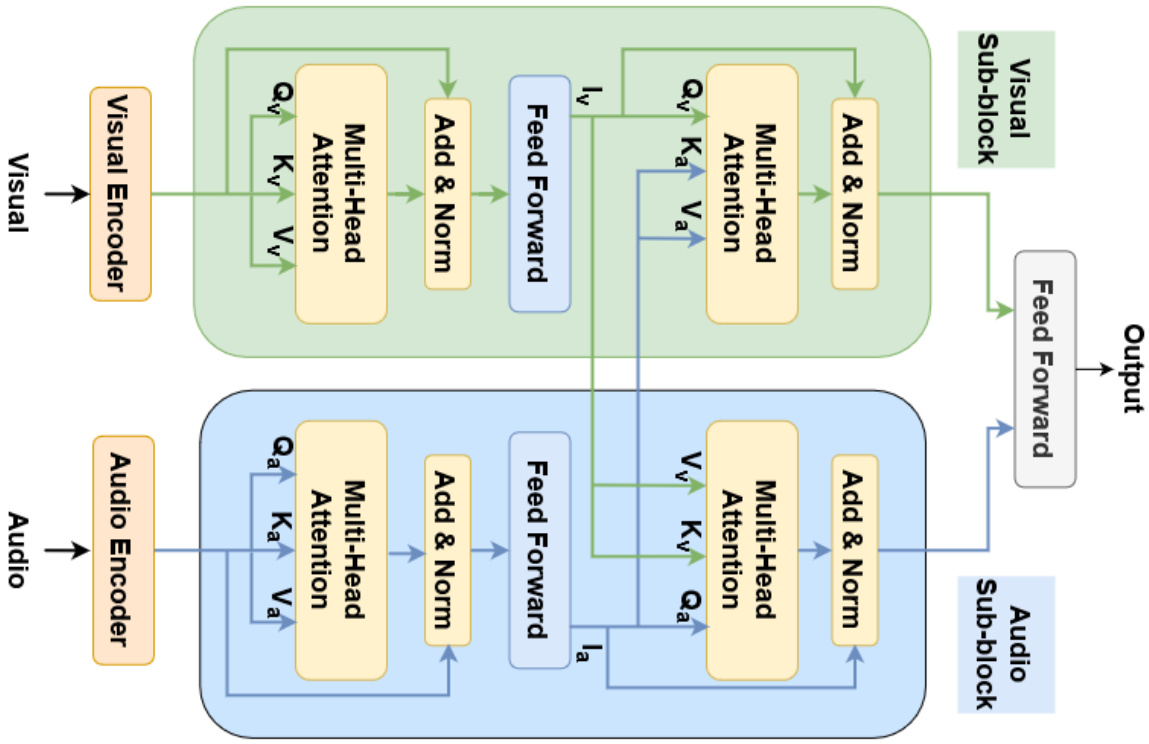


Figure 2.5: The co-attention block.

(Norm) and a residual connection are applied after each attention module. A fully-connected feed-forward network (FC) is also added.

In the visual sub-block, the extracted visual embedding from the visual encoder is first fed into a multi-head self-attention module to get the intermediate visual representation, I_v, embedding important visual information. Similarly, we can get the intermediate audio representation, I_a, in the audio sub-block. Specifically, I_v and I_a can be computed as follows.

$$\begin{aligned} I_{iv} &= FC(Norm(MultiHead(z_{iv}, z_{iv}, z_{iv})) + z_{iv}) \\ I_{ia} &= FC(Norm(MultiHead(z_{ia}, z_{ia}, z_{ia})) + z_{ia}) \end{aligned} \tag{2.3}$$

where z_{iv} and z_{ia} denote the output features from the visual encoder and the audio encoder for segment i, respectively.

Next, in the visual sub-block, I_{ia} as key and value and I_{iv} as query are passed into the multi-head co-attention module. In this way, we can enforce the visual sub-block to focus on the information related to audio. Similarly, in the audio sub-block, we feed I_{iv} as key and value and I_{ia} as query into the second multi-head attention layer. Hence, the final output features of vision and audio, F_{iv} and F_{ia}, can be computed as:

$$\begin{aligned} F_{iv} &= FC(Norm(MultiHead(I_{iv}, I_{ia}, I_{ia})) + I_{iv}) \\ F_{ia} &= FC(Norm(MultiHead(I_{ia}, I_{iv}, I_{iv})) + I_{ia}) \end{aligned} \tag{2.4}$$

Thus, the audio sub-block tends to focus on the information corresponding to vision. Consequently, two sub-blocks can find important information about themselves, as well as their relationships. Using such a mechanism, we capture the interaction between visual and audio components. Finally, we combine the guided visual representation and the guided audio representation by applying an FC layer as:

$$F_i = FC(concat(F_{iv}, F_{ia})) \tag{2.5}$$

Therefore, the final output of this co-attention block is the joint representation of vision and audio for each segment i.

Sec2Sec Structure. To capture the temporal information in the video clip sequence, we feed the joint representation of each segment, F_i, from the co-attention block to an LSTM network, illustrated in Figure 2.6. The LSTM network is defined as follows:

$$
\begin{aligned}
u_i &= \sigma(W_{Fu}F_i + W_{hu}h_{i-1} + b_u) \\
f_i &= \sigma(W_{Ff}F_i + W_{hf}h_{i-1} + b_f) \\
o_i &= \sigma(W_{Fo}F_i + W_{ho}h_{i-1} + b_o) \\
\tilde{c}_i &= tanh(W_{Fc}F_i + W_{hc}h_{i-1} + b_c) \\
c_i &= f_i \odot c_{i-1} + u_i \odot \tilde{c}_i \\
h_i &= o_i \odot tanh(c_i)
\end{aligned}
\tag{2.6}
$$

where $\sigma(\cdot)$ is an activation function. $\odot$ denotes the Hadamard product. W and b are weights and biases to be learned during training. h_i denotes the hidden state at step i. u_i, f_i, o_i and c_i denote the update gate, forget gate, output gate and cell gate, respectively.

Predictor. We apply an FC along with a sigmoid function to the output from the LSTM network at the last step to make emotion predictions.

2.5 Experiments

2.5.1 Dataset

We use a large-scale publicly available dataset LIRIS-ACCEDE [10] to evaluate the effectiveness of our proposed model. The dataset contains $9,800$ videos extracted from 160 films. Most films are from the popular video-sharing platform VODO. The languages spoken in the films are mainly English with a small set of 9 other languages subtitled in English. There are also 14 silent movies. The films cover 9 main categories of movies including action, comedy, drama, etc. The video clips last between 8 seconds and 15 seconds. To annotate each video, researchers

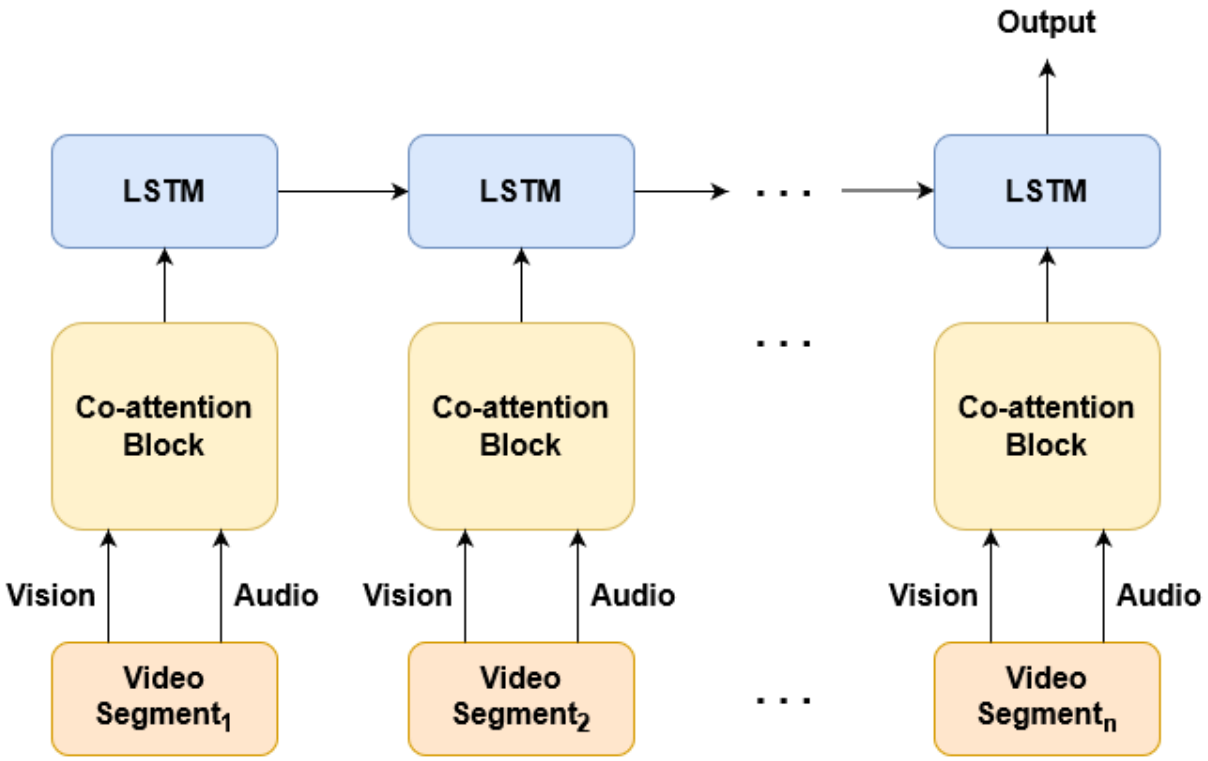


Figure 2.6: The Sec2Sec Structure.

recruited $1,517$ annotators from 89 countries to minimize the cultural impact on video emotions. After watching each video, each annotator provides ratings for valence and arousal in the range between 1 and 9, respectively. The annotation of each video is calculated by taking an average of all the individual annotations. In this paper, we adopt a binary classification approach based on the existing literature [1], where the threshold used to separate high from low arousal (or valence) is 5. The dataset is split into 70-10-20% for training, validating and testing, respectively. For evaluation, we adopt two standard metrics for emotion classification tasks (valence and arousal), accuracy and F_1 score.

2.5.2 Implementation Details

We train all models on four NVIDIA GeForce 3090 24GB GPUs with 250 epochs[2]. Our model is trained to minimize the binary cross entropy loss with the Adam optimizer [45]. We set up an early stopping mechanism, where the training stops if the validation loss increases for 5 consecutive epochs. We use the grid search strategy to find a relatively optimal set of hyperparameters. Specifically, the learning rate, batch size, and the number of heads are searched within ranges of [1e-8, 1e-5], [8, 16, 32, 64, 128], and [8, 16, 32, 64, 128], respectively. In each experiment, we use the model with the best validation accuracy to report results on the holdout testing set.

Table 2.3: Performance comparison of our model with baselines for arousal prediction.

	Method	Modality	Accuracy	F_1 Score	Avg Training Time Per Epoch (min)
Baselines	ViT[31]	Audio	0.7823	0.8768	1:55
	ViViT [9]	Vision	0.7853	0.8795	1:32
	CMA [22]	Audio and Vision	0.5680	0.6768	4:50
	AVM [47]	Audio and Vision	0.7756	0.8722	4:49
	ViT-ViViT	Audio and Vision	0.7517	0.8541	3:31
	Co-attention	Audio and Vision	0.5599	0.6603	4:50
Variants	Sec2Sec Audio	Audio	0.7832	0.8780	1:51
	Sec2Sec Vision	Vision	0.7766	0.8733	0:20
	Sec2Sec SA-SA	Audio and Vision	**0.7990**	**0.8876**	2:17
	Sec2Sec SA-CA	Audio and Vision	**0.7949**	**0.8840**	2:14

2.5.3 Baselines

We evaluate the performance of our proposed model (called **Sec2Sec SA-CA**) with several state-of-the-art methods.

[2]For the purpose of reproducibility, our implementation is publicly available at `https://github.com/nestor-sun/sec2sec`.

Table 2.4: Performance comparison of our model with baselines for valence prediction.

	Method	Modality	Accuracy	F_1 Score	Avg Training Time Per Epoch (min)
Baselines	ViT[31]	Audio	0.7022	0.8154	1:52
	ViViT[9]	Vision	0.7002	0.8234	1:29
	CMA [22]	Audio and Vision	0.6078	0.7033	6:05
	AVM [47]	Audio and Vision	0.7205	0.8287	4:49
	ViT-ViViT	Audio and Vision	0.69	0.8009	3:32
	Co-attention	Audio and Vision	0.5864	0.6688	4:50
Variants	Sec2Sec Audio	Audio	0.7021	0.8191	1:50
	Sec2Sec Vision	Vision	0.6970	0.8179	0:20
	Sec2Sec SA-SA	Audio and Vision	0.7047	0.8179	2:14
	Sec2Sec SA-CA	Audio and Vision	**0.7322**	**0.8372**	2:15

CMA [22]: A cross-modal attention (CMA) Transformer network developed for audio-visual correspondence prediction. We train a CMA using audio and vision.

AVM [47]: A bi-modal (audio and vision) deep network, consisting of R(2+1)D ResNet and ResNet networks, was developed for emotion prediction. Specifically, a pre-trained R(2+1)D ResNet is used to extract visual features, and audio features are extracted using ResNet. Lastly, a fully-connected feed-forward network is added to fuse two types of features for prediction.

ViT [31]: Vision Transformer (ViT) is a Transformer-based model for image classification, which have been demonstrated outstanding performance over convolutional neural networks. We fine-tune a pre-trained ViT using the audio component (i.e., treated as images), since ViT can only process individual images rather than a sequence of images.

ViViT [9]: Video Vision Transformer (ViViT) is a Transformer-based model, designed for video classification. It can capture spatio-temporal information. We train a ViVit using vision, since ViViT can process a sequence of images.

ViT-ViViT: We implement a bi-modal (audio and vision) network by combining ViT and ViViT

to extract audio and visual features, respectively. Two types of features are concatenated and fed into a fully-connected feed-forward network for emotion prediction.

We also add a co-attention network as another baseline and 3 variants of our model for comparison to understand the role of each design in our model (e.g.,. uni vs. bi-modal, co-attention).

Co-attention: It only trains a multi-head co-attention model without segmenting each video into audio and vision components.

Sec2Sec Audio: We use a multi-head attention model with 2 layers of self-attention that only relies on the audio component. Specifically, we first divide each audio into n segments. Each segment goes through a pre-trained ResNet as an audio encoder. The output from the audio encoder for each segment is sent to two multi-head self-attention layers. The output for each segment is then fed into an LSTM network. Finally, a fully-connected layer is added to predict emotions.

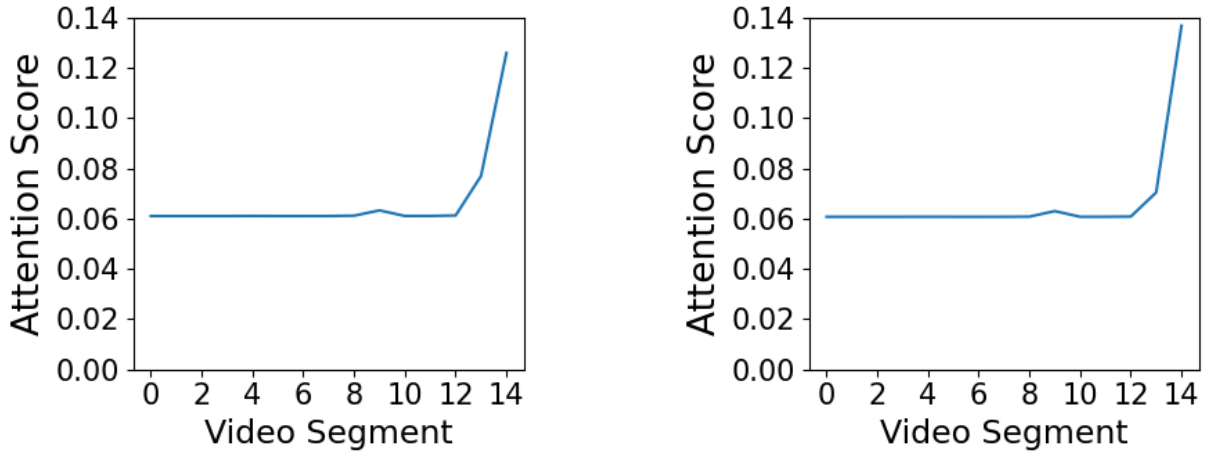
Sec2Sec Vision: It is a multi-head attention model with 2 layers of attention using only the visual component. Similar to Sec2Sec Audio, we first split each video into n visual segments. Each visual segment is fed into the visual encoder. And the corresponding output will be sent to two multi-head self-attention layers. Lastly, an LSTM and a fully-connected layer are applied to predict emotions.

Sec2Sec SA-SA: It is a multi-head attention model with 2 attention layers using both audio and vision. Unlike the Sec2Sec SA-CA model, which uses one self-attention layer and one co-attention layer, we design a variant (called Sec2Sec SA-SA) that uses two self-attention layers to capture the intra-modal dependencies within segments.

2.5.4 Results

Overall performance Table 2.3 and 2.4 present the experimental results for arousal and valence, respectively. Our proposed Sec2Sec models achieve the best performance on two evaluation metrics for arousal prediction. They surpass three bi-modal (audio and vision) methods and the co-attention approach in terms of accuracy and efficiency, demonstrating the benefit of incorporating LSTM (Sec2Sec) into the video understanding framework. They also outperform Sec2Sec Audio and Sec2Sec Visual, indicating that using both audio and visual components is more effective than using either modality alone. Moreover, Sec2Sec SA-SA and Sec2Sec SA-CA obtain comparable results, suggesting that the interaction between audio and visual features is not essential for predicting arousal. It is noteworthy that ViT-ViViT performs worse than ViT and ViViT, indicating that a single fully-connected layer fails to adequately capture the interaction between audio embeddings and visual embeddings that are derived from ViT and ViViT. We have similar observations for valence prediction, in terms of performance comparison with baselines.

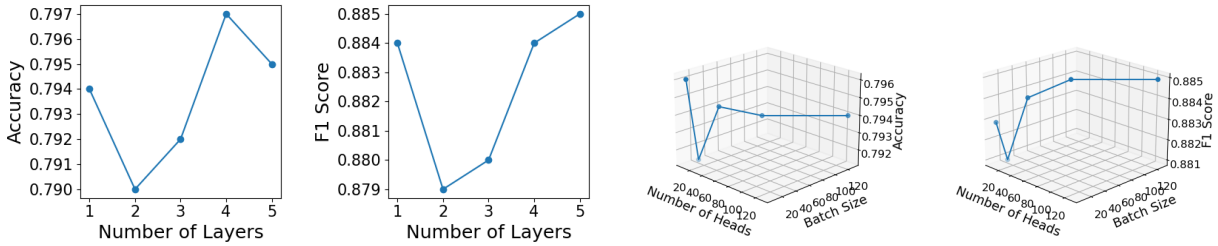
Model Interpretability we now turn to assess the contribution of each video segment (i.e., every one-second clip) to emotion prediction. To do so, we modify the Sec2Sec structure by substituting LSTM with an attention-based LSTM proposed by [100]. We adopt the identical hyperparameters as the Sec2Sec Co-attention model. After training, we obtain the learned LSTM attention values and normalize the values by applying a softmax function. The attention values of each video segment for valence and arousal are plotted in Figure 2.7a and 2.7b, respectively. Similar patterns are observed in both figures. They tell that the emotion prediction power reaches the highest for the last 3 seconds of the video. It may suggest that emotions are mostly influenced by last 3 seconds. Moreover, the impact of video segments increases as they approach the end of a video.



(a) Valence

(b) Arousal

Figure 2.7: The LSTM Attention.



(a) Accuracy regarding # of LSTM layers.

(b) $F1$ regarding # of LSTM layers.

(c) Accuracy regarding # of heads and batch size.

(d) $F1$ regarding # of heads and batch size.

Figure 2.8: Effect of key hyperparameters of Sec2Sec SA-CA on accuracy and $F1$ score.

We hypothesize that when annotators rate each video, their decisions are dominated by the last 3 seconds of each video, which aligns well with the theory of recency bias in psychology [61].

2.6 Ablation Study

To assess the impact of several key hyperparameters on the model performance, we conduct several additional experiments. Results are shown in Figure 2.8. Due to the space limitation, we only report the results for arousal while valence has similar patterns.

Impact of the number of heads and batch size. We examine the model performance by varying

Table 2.5: Comparison of batch vs. layer normalization.

Method	Emotion	Accuracy	$F1$ score
layer	arousal	**0.7944**	**0.8840**
batch	arousal	0.7924	0.8832
layer	valence	**0.7271**	**0.8361**
batch	valence	0.7220	0.8322

the number of heads and batch size simultaneously from 8 to 128. The model achieved the best accuracy when both the number of heads and batch size were set to 8, while the best $F1$ score was achieved when both were set to 64 or 128.

Impact of the number of LSTM layers. We vary the number of LSTM layers from 1 to 5. We find that even the model with 4 LSTM layers achieves the best accuracy, $F1$ score. is not significantly different with 1 LSTM layer, 4 LSTM layers or 5 LSTM layers. It is worth noting that more LSTM layers can increase memory consumption if batch size remains the same.

Layer vs. batch normalization. To examine the effect of normalization methods on perceived emotion recognition, we contrast layer normalization with batch normalization, which are commonly used in Transformers and computer vision models, respectively [42]. As Table 2.5 shows, layer normalization outperforms batch normalization for both arousal and valence predictions on accuracy and $F1$.

2.7 Conclusion

In this study, we propose a novel Sec2Sec Co-attention Transformer model for perceived emotion classification, which leverages self-attention and co-attention mechanisms to encode and fuse multimodal features. We have evaluated our model on the LIRIS-ACCEDE dataset and achieved

better results compared with state-of-the-art baseline approaches. The results show the effectiveness of our Sec2Sec structure and the importance of inter-modal interaction for emotion prediction. We also introduced an attention-based LSTM mechanism to explore the contribution of each second clip of a video to the overall emotion prediction. Our work has several implications for multimodal emotion recognition research and applications. First, it demonstrates that Sec2Sec models can improve both performance and efficiency over traditional encoder-decoder models. Second, it reveals that co-attention can capture rich inter-modal relations that are essential for emotion prediction. Third, it provides a novel way to interpret the model prediction by visualizing the attention weights over video segments. Future work can extend our model to other multimodal tasks such as video sentiment analysis and audio-visual alignment analysis.

Chapter 3: Multimodal Co-attention Transformers for Video-Based Apparent Personality Understanding

3.1 Introduction

In the dynamic landscape of communication and media consumption, video content has emerged as a dominant and influential medium. For example, in 2023, video content makes up for 82% of the Internet traffic [79]. Video content is not essential at a macro level, but also at a micro level. Many social media platforms, such as TikTok and Instagram, have started providing video-sharing services, which plays a critical role in people's daily life. For instance, More than 78% of viewers consume video content every week, and 55% of them engage every day [79]. In addition, 93% of companies acquire new customers via social media videos [105]. In light of the burgeoning prevalence of video content, it becomes imperative to comprehend the personalities embodied by the presenter or influencer featured in each video. The elucidation of such personalities holds substantial potential for enhancing the efficacy of subsequent predictive analytics. The personality traits of presenters or influencers can serve as robust predictors, thereby contributing to more accurate forecasts in downstream predictive applications. Thus, a thorough understanding of these personalities is not just beneficial, but essential for leveraging the full potential of predictive analytics in the realm of video content.

Figure 3.1: Images sampled from videos in the First Impressions dataset that exhibit varying degrees of personality traits.

(a) High O (b) High C (c) High E (d) High A (e) High N

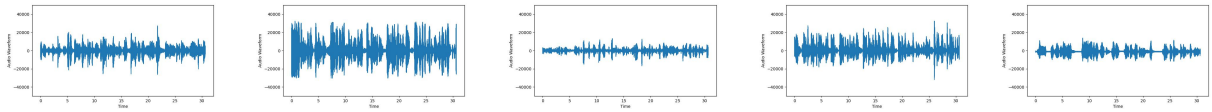
(f) Low O (g) Low C (h) Low E (i) Low A (j) Low N

Note: These traits are represented by the acronym OCEAN, where O stands for Openness, C for Conscientiousness, E for Extraversion, A for Agreeableness, and N for Neuroticism. High personality levels are often recognized by exhibiting a friendly face with a bright background, while low personality levels are recognized by showing an unhappy expression with a dark background.

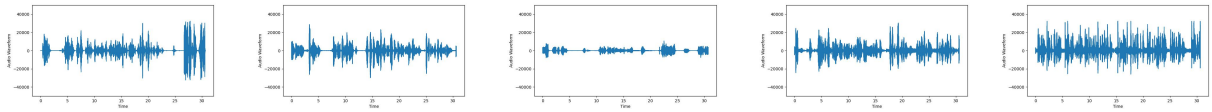
Personality plays a pivotal role in shaping human interactions, decision-making processes, and overall behavioral patterns [6]. For instance, product recommendations and the effectiveness of word-of-mouth are largely affected by personality in digital marketing [2]. In human resources, personality can help predict a candidate's suitability for a specific job [49]. In addition, [62] found that a CEO's personality plays an important role in driving a company's strategic flexibility. In the context of Information Systems, [28] find a significant relationship between personalities and technology acceptance and adoption. These studies emphasize a relationship between personalities and downstream outcomes.

Given its impact, the study of video-based personality detection holds tremendous potential across various disciplines, such as psychology, marketing, human-computer interaction, and social sciences. By discerning the personalities projected through videos, researchers and practitioners can gain valuable insights into how individuals are perceived by others, the effectiveness of

Figure 3.2: Audio waveforms in the First Impressions dataset that exhibit varying degrees of personality traits.



(a) High O (b) High C (c) High E (d) High A (e) High N



(f) Low O (g) Low C (h) Low E (i) Low A (j) Low N

Note: These traits are represented by the acronym OCEAN, where O stands for Openness, C for Conscientiousness, E for Extraversion, A for Agreeableness, and N for Neuroticism. Individuals with high personality levels often have a high voice when speaking, whereas individuals with low personality levels have a low voice.

persuasive communication strategies, and the influence of personality on audience engagement [2, 49, 72]. Moreover, as social media platforms increase and user-generated video content becomes increasingly prevalent, video-based personality detection becomes a valuable tool for understanding broader societal trends, cultural influences, and collective attitudes.

The recent surge in interest in video-based apparent personality trait prediction has underscored several non-trivial challenges, primarily due to the unique characteristics inherent to the video-based personality setting. Firstly, a video typically comprises three types of information: visual, auditory, and textual. Each of these modalities may contain crucial information that could significantly enhance the accuracy of predictions. Research has demonstrated that personality traits are evident in appearance, expression, voice [51, 63, 76]. For the purpose of illustration, Figures 3.1 and 3.2 depict distinct visual and acoustic patterns corresponding to different levels of the Big-Five personality traits. Secondly, [44] posits that capturing modality interactions is essential for making accurate predictions. This is further corroborated by the McGurk effect [58],

which underscores the importance of the interaction between auditory and visual modalities. We contend that capturing the interactions among all three modalities is crucial as a good alignment can synergistically aid audiences in better inferring personalities, while a poor alignment can adversely affect the perceptions of personalities. Thirdly, existing approaches often necessitate high-resolution images, typically in the dimension of 224×224, which can be computationally expensive. Finally, deep learning models are often criticized for being 'black boxes' due to their lack of interpretability. However, in the context of video-based perceived personality settings, it is essential to provide interpretability that offers practical implications for both presenters or influencers and researchers or platforms. This interpretability not only demystifies the underlying mechanisms but also facilitates more informed decision-making processes.

Several models have been proposed. For example, [102] proposes a bi-modal network to process visual and audio information and predict personality in video. A tri-modal network is also processed to predict video personality by taking in visual, audio and text information [83] However, most of these works train a different model for each modality independently and combine predictions using ensemble methods such as taking an average of the predictions generated by different modalities. More importantly, all the models mentioned above require high-resolution pictures (e.g. 224×224) in order to perform well. Processing high-resolution pictures is computationally expensive.

To solve these limitations and improve prediction accuracy, in this paper, we propose a Multimodal Co-attention network based on the multi-head self-attention mechanism proposed in Transformer [97]. Specifically, we develop a visual encoder extended from [31] along with a newly proposed hierarchical positional encoding mechanism to efficiently extract visual features, two linear regressors to extract audio and text features. We further develop a Multimodal Co-

attention Transformer to understand the complex interactions among visual, audio and text components efficiently.

To evaluate our work, we conduct experiments on a real-world dataset, First Impressions, with 10,000 videos to demonstrate the usefulness and value of the proposed model. The experimental results not only show that the proposed model outperforms seven state-of-the-art baselines, but also has improved the computational costs. Furthermore, we conduct a series of interpretability analyses to demonstrate the model's decision process. Our analysis uncovers the useful factors that can be used to predict personality traits at modality-, vision- and image- level, which can serve as a guideline for presenters, influencers and platforms that seek to improve perceived personality.

To conduct our interpretability analysis, we calculate the contributions of inputs by computing the Integrated Gradient for each input proposed by [88]. For the modality-level interpretability analysis, the contributions of the inputs are aggregated into three modalities: audio, vision and text. Our results show that text information is less important than the other two modalities. The relative importance of audio and vision depends on the specific personality traits. For agreeableness, neuroticism and openness, audio is more important than vision. For extraversion, vision is more important than audio. For conscientiousness, audio and vision are equally important. For the vision-level interpretability analysis, the contributions of visual inputs are aggregated by the time point of each input image, which enables us to investigate at which time point an image is more important at predicting personality traits. For agreeableness, extraversion, neuroticism and openness, the importance of images decreases over time, indicating the first impression does matter when perceiving personality. Interestingly, for conscientiousness, the importance of images increases over time, suggesting . For the image-level analysis, the contributions of visual

inputs are aggregated into different regions of an image, which allows to study the importance of different regions. The results show that hand movements and background are more important than faces at predicting personalities. We believe the findings from the interpretability analyses serve as guidelines for influencers or presenters to better design and create video content and for audiences to infer personalities.

In addition to the interpretability analysis, we use a real-world case study to showcase the usefulness of the proposed model. Specifically, we collected the MBA admission data from a major university in the United States. In the application process, each applicant is required to record a up-to-one-minute video to answer a specific question. The staff uses the video as evidence to evaluate the communication skills and English proficiency of each candidate. We utilize our model to generate the five personality predictions for each candidate, and the impact of the predicted personality traits as a persuasion tool on the admission outcome. Our results show that candidates perceived as conscientious, extroverted and agreeable are associated with a higher chance of being admitted. Among these three traits, being perceived as agreeable has an even higher probability of being admitted.

In summary, this study makes the following contributions. First, we introduce a Multimodal Co-attention network, coupled with a novel hierarchical positional encoding mechanism. This sophisticated architecture adeptly processes information from visual, acoustic, and textual modalities. Impressively, our approach outperforms state-of-the-art baselines, showcasing its remarkable performance. Second, we validate our proposed model in extracting valuable insights from low-resolution (64×64) images. Notably, even with a compact latent representation space of just 512 dimensions, our model excels. Moreover, it achieves this while demanding minimal training time. Third, we rigorously conduct interpretability analyses, shedding light on the intricate decision-

making process of our method. This offers a deeper understanding of its working mechanism, enhancing its transparency. Finally, to underscore the practical utility of video-based personality detection, we present a compelling case study. This aptly demonstrates the model's efficacy in predicting MBA admission outcomes, showcasing its real-world application.

The remainder of the paper is organized as follows. In Section 2, we discuss prior work on personality prediction, Transformer in computer vision as well as recent work on multimodal learning. In Section 3, we give a brief overview of Transformer, followed by the proposed modeling in Section 4. In Section 5, we present the experimental details. Section 6 presents the results. We conclude our study in Section 7.

3.2 Related Work

3.2.1 Personality

Personality is often defined as a distinct combination of cognitive, affective, and behavioral traits [6]. It is considered to be relatively stable compared to emotion [24]. In the business-related literature, the Big-Five personality traits (OCEAN: **O**penness, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**euroticism) are widely used to describe a person's personality [30]. They are defined as follows [35]:

- Openness emphasizes imagination and insight.
- Conscientiousness denotes organization and responsibility.
- Extraversion represents sociability and energy.
- Agreeableness reflects compassion and trust.

- Neuroticism involves anxiety and depression tendencies.

The influence of personality traits on various aspects of life and decision-making processes has been extensively studied in the literature. For instance, the personalities of senior chief executive officers (CEOs) have been found to significantly correlate with their companies' financial outcomes, such as cash holdings, investment, and interest coverage [109]. In particular, conscientiousness has been negatively associated with a company's strategic flexibility, while agreeableness, extraversion, and openness have shown positive associations [62]. The relationship between personality traits and online shopping behaviors has also been explored. Consumers with higher degrees of neuroticism, agreeableness, or openness tend to be utility-motivated to shop online [96]. Furthermore, hedonic purchase motivation is positively influenced by neuroticism, extraversion, and openness [96]. In the realm of information-seeking tasks, individuals high in conscientiousness performed fastest, followed by those high in agreeableness and extraversion [4]. Moreover, personality traits have been linked to social media usage and engagement. Specifically, openness and extraversion are the two most significant positive predictors of social media use. Conscientiousness, agreeableness, and neuroticism were also considered important but to a lesser degree [48]. These studies underscore the crucial relationship between personality traits and decision-making choices across various domains.

One of the biggest limitations of these studies is that how personality traits are derived. The majority of literature requires the completion of long questionnaires to determine personality traits [26], which is time-consuming and burdensome. With the recent data explosion in user-generated content, such as in social media, it has become almost impossible to conduct questionnaire-based research. A lot of effort has been devoted to automatic personality detection from text.

Early works include using word count to classify personalities. For instance, [2] use Linguistic Inquiry and Word Count (LIWC) to classify text personalities. Recently, research has started adopting deep learning techniques to predict personality traits. One advantage of employing deep learning methods is their ability to learn word embeddings that capture rich contextual information in text, facilitating the learning of document-level representations by the models. For instance, a deep convolutional neural network (CNN) has been developed to predict personality from text information and has been demonstrated to outperform traditional machine learning techniques [86]. [111] found that CNNs outperform recurrent neural networks (RNNs), such as long short-term memory (LSTM) and gated recurrent units (GRU), in predicting personality. Attention techniques have also been incorporated into CNNs to enhance their performance. For example, word-level attention has been proposed to learn document-level semantic features [108], while message-level attention has been employed to leverage the relative weight of users' social media posts, yielding impressive results [57]. In addition, [37] utilize three pre-trained language models, BERT, RoBERTa and XLNet, to predict text personalities by averaging predictions. [109] develop a hierarchical attention network to classify text personalities.

However, even though video content has been booming, there are only a few studies that focus on video-based personality detection. In the next section, we will discuss current video-based personality efforts as well as their limitations.

3.2.2 Video-Based Deep Personality Prediction

Personality prediction has recently emerged as a popular research area, with a focus on utilizing deep learning techniques to predict personality traits from unstructured data sources

such as text and video. There is little research has focused on predicting personality traits from user-posted social media videos. These videos typically consist of at least two modalities: vision and audio, with some also including text. Various methods have been proposed to process and combine visual and audio data. Most existing video personality prediction models extract information and make predictions from each data source (vision, audio or text) separately and then employ ensemble methods, such as averaging, to combine predictions. For instance, [102] developed a Descriptor Aggregation network to predict personality traits from video-sampled images and a linear regressor to predict personality traits from audio, averaging the predictions from these two models to make final predictions. [83] utilized pre-trained VGG-16 and ResNet models to predict personality from audio and images respectively, a linear regressor to predict personality from text, and averaged the predictions from all three models to make final predictions. Another approach involves extracting features from each source and using a fully connected feed-forward network to fuse embeddings from two or three modalities. [82] proposed two techniques for predicting video personality traits: one using a 3D convolution network to extract visual features and a linear regressor to extract audio features, with a fully connected network fusing the two modalities to make predictions; the other splitting a video into several equal-length parts and using a linear regressor and CNN to extract audio and visual features respectively for each part, with a fully-connected feed-forward network combining the embeddings as the latent representation for that part before entering an LSTM network sequentially to make final predictions. [36] developed two CNNs to extract audio and visual features and employed a fully-connected network to combine the embeddings and make predictions. However, the majority of these models fail to capture the interactions between audio and vision, which is crucial for multimodal learning [44]. More importantly, these models require high-resolution pictures (e.g.,

224×224) to process images, which is computationally expensive.

3.2.3 Transformers in Computer Vision

The Transformer model, initially proposed for machine translation tasks in the realm of natural language processing (NLP) [97], has seen a surge of interest for its application in computer vision (CV) tasks, positioning it as an alternative to convolutional neural networks (CNNs). Several studies have made significant strides in this area. For instance, the Vision Transformer (ViT) [31] applies a Transformer model to linearly projected sequences of image patches for full image classification. The Swin Transformer enhances the ViT by introducing a hierarchical Transformer architecture coupled with a shifted window scheme [53]. These models serve as two representative Transformer-based models for image classification tasks. In the context of video classification tasks, the Video Vision Transformer (ViViT) extends the ViT by proposing two methods for embedding video samples: uniform frames sampling and tubelet embedding. It also introduces four model variants based on the Transformer: spatiotemporal attention, factorized encoder-decoder, factorized self-attention, and factorized dot product attention [9]. The Video Swin Transformer further extends the Swin Transformer by introducing a 3D-shifted window-based multi-head self-attention module and a locality inductive bias to the self-attention module [54]. Since the advent of pure Transformer-based models for computer vision tasks, they have been adopted for a diverse range of applications, including semantic segmentation [113], action recognition [14], and object detection [16]. This underscores the versatility and efficacy of Transformer models across various domains.

3.2.4 Multimodal Learning

Multimodal learning, a deep learning technique, involves the assimilation of information from diverse modalities such as images, text, audio, and video. Given the inherent multimodal nature of videos, a substantial body of literature has focused on learning a joint representation of audio and vision to predict audio-visual synchronization. For instance, [8] trained an audio-visual cross-modal network from scratch to predict video correspondence. [7] utilized one clustered modality as a supervisory signal for another modality and predicted correspondence between two modalities. [22] further developed three self-supervised co-attention-based networks to discriminate visual events related to audio events. However, only a handful of research studies have concentrated on handling vision, audio, and text. For example, [5] applies contrastive learning to vision, audio, and text to learn video-level representations for self-supervised learning tasks. Most multimodal learning models employ a standard Transformer as the backbone network to learn the interactions among different modalities. For instance, VATT [3] proposes a Transformer-based self-supervised learning model that can process audio, visual, and text information and uses a standard Transformer model as the backbone network. [34] develops an omnivore network that uses a standard Transformer network to learn representations from images, videos and 3D View data. While it is relatively straightforward to train a standard Transformer in terms of implementation, the computational cost can escalate when the latent representation space enlarges. Our method aims to enrich the multimodal learning literature by proposing a Multimodal Co-attention Transformer that can efficiently process information from three different modalities.

3.3 Preliminaries

In this section, we provide a brief explanation of the functionality of multi-head self-attention. The key component of the Transformer [97] maps a query vector Q, a key vector K and a value vector V to an output (embedding), as demonstrated in Equation 3.1.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{3.1}$$

Self-attention is achieved by computing the similarity scores between the query matrix Q and the key matrix K using the scaled dot product. This is obtained by dividing the dot product of Q and K by the square root of the dimension of Q, denoted by d. These scores are then converted into probabilities by applying a softmax function. The resulting probabilities are used to weight the values in the value matrix V, producing the final output for the next layer. The underlying principle of this mechanism is to assign greater importance to vectors with higher probabilities in V in subsequent layers.

Despite its effectiveness, a single self-attention layer may constrain a model's capacity to attend to multiple positions simultaneously without sacrificing attention to other positions. A multi-head attention mechanism is introduced to address this limitation, which has been shown to enhance overall model performance. This mechanism comprises h parallel self-attention sub-layers, referred to as 'heads', each of which learns distinct query, key, and value matrices. These heads project input features into different subspaces, and their output features are concatenated

into a single matrix for subsequent processing by downstream layers, as shown in Equation 3.2.

$$MultiHead(Q, K, V) = Concat(h_1, h_2, ..., h_h)W^O$$
$$where\ h_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3.2)$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_{model}}$ are learnable weights for each head i. $W^O \in \mathbb{R}^{h \times d_{model} \times d_{model}}$ is the projection weight.

A fully-connected feed-forward network is applied after the self-attention layer. In addition, a residual connection and a normalization function are applied to each sub-layer.

3.4 Our Model

In this section, we introduce our proposed model, which is designed to extract and analyze visual, acoustic, and textual representations, as well as the interactions among these three modalities. As illustrated in Figure 4.2, our model comprises three primary components:

1. **Encoder Network**: The encoder network is composed of visual, audio, and text encoders that are responsible for extracting the respective features from each modality.

2. **Multimodal Co-attention Transformer Network**: This network is designed to capture the interactions among the three modalities through the use of a multimodal co-attention mechanism.

3. **Predictor**: The output from the Multimodal Co-attention Transformer Network is fed into a fully connected feed-forward network, which generates predictions regarding personality traits.

Visual encoder and Hierarchical Positional Encoding. For the visual encoder, we build

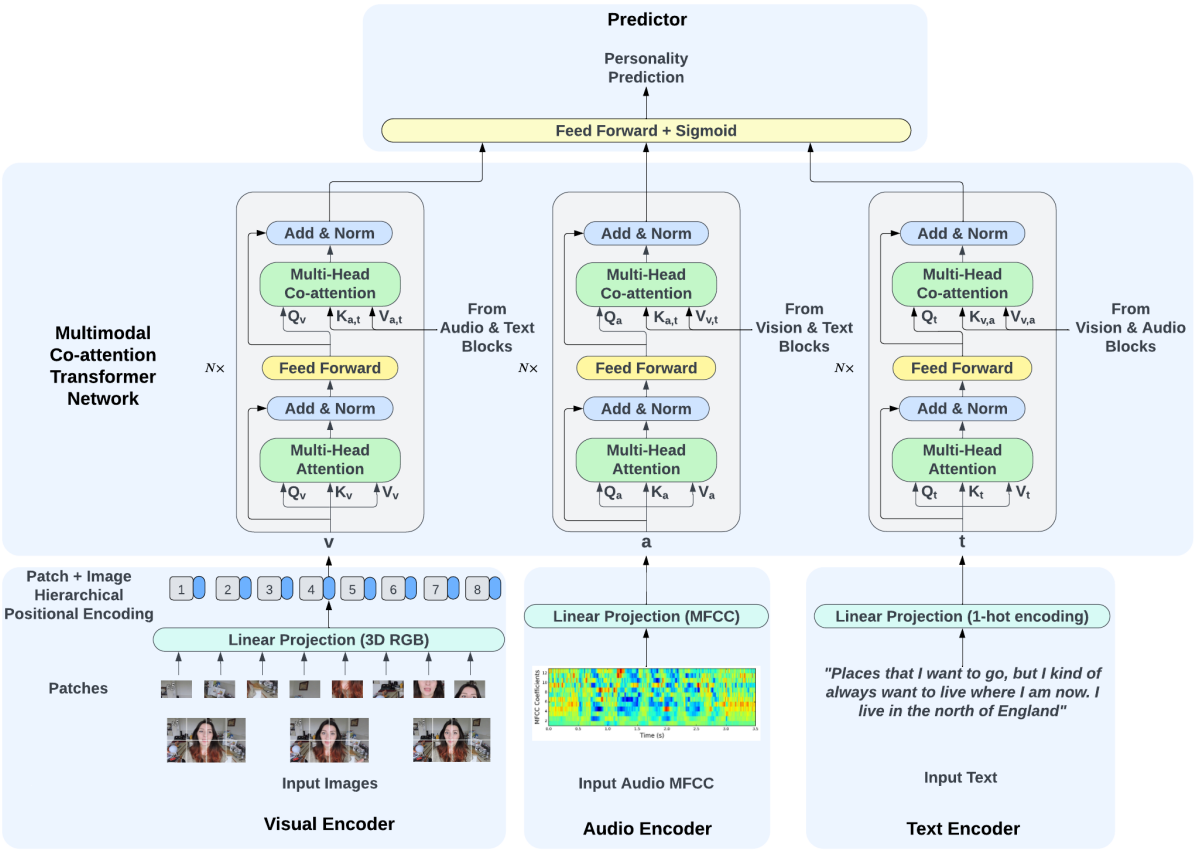


Figure 3.3: An overview of our proposed model: Multimodal Co-attention Transformer.

upon the work of the Vision Transformer (ViT) [31]. Our visual encoder takes as input a 3-channel Red-Green-Blue (RGB) representation of n sampled image frames, with a size of $[3, H, W]$. Each image is partitioned into patches of size $[h, w]$, resulting in a total of $[H/h] \times [W/w]$ patches. Additionally, we propose a hierarchical positional encoding mechanism to incorporate positional information into the model. In this study, we utilized a sample size of 100 images per video. Contrary to the majority of studies that resize images to a higher resolution of $[224, 224]$, we opted for a lower resolution size of $[64, 64]$ for each image. The patch size is a hyperparameter that requires tuning.

In the visual encoder, which lacks both recurrence and convolution, a hierarchical positional

encoding method is introduced to encode the position of each patch within each sampled image and the position of each image within each video. This enables the model to comprehend the position of each patch or image. The positional encodings of patches and images share the same dimension as that of each patch, allowing for the addition of encodings and embeddings. Specifically, the hierarchical positional encoding method encompasses two components: patch positional encoding and image positional encoding. Both encodings possess dimensions equivalent to those of individual patches, facilitating their effortless integration with patch embeddings. Additionally, in the visual encoder, which lacks both recurrence and convolution, we introduce a hierarchical positional encoding method to encode the position of each patch within each sampled image and the position of each image within each video, which enables the model to comprehend the position of each patch or image. Specifically, the hierarchical positional encoding method encompasses two components: patch positional encoding and image positional encoding. Both encodings possess dimensions equivalent to those of individual patches, facilitating their effortless integration with patch embeddings. The positional encodings of patches and images share the same dimension as that of each patch, allowing for the addition of encodings and embeddings. We build upon the positional encoding, $PE_{pos,i}$, proposed in [97], which uses sine and cosine encoding functions written as:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d}) \tag{3.3}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d}) \tag{3.4}$$

where pos is the position and i is the dimension.

Therefore, we get the positional encodings for patch p in each image and image m in each

video, $PE_{p,i}$ and $PE_{m,i}$, based on the equation above. Together, we get the hierarchical positional encoding for patch p in image m in a video as

$$PE_{p,m} = PE_{pos_{p,i}} + PE_{pos_{m,i}} \tag{3.5}$$

After injecting positional encoding, each patch is linearly projected to a latent representation with a dimension of l. The latent representation for each patch is then concatenated to form the visual embeddings with a dimension of $l \times [H/h] \times [W/w]$.

Audio encoder. In the audio modality, our approach involves several stages. First, we extract the raw audio signal from each video. Subsequently, all audio signals are re-sampled from their original rate of 44.1kHz to a standard rate of 16kHz. From these re-sampled signals, we extract 2-dimensional Mel-Frequency Cepstral Coefficients (MFCC) [27], which are designed to model how the human ear perceives and distinguishes between different sound frequencies [71]. These 2-dimensional MFCCs are then flattened into 1-dimensional representations for input into the audio encoder. We conducted experiments with various audio features, including log bank filters and raw audio waveforms, among others. Our results indicated that MFCCs provided the best performance, and thus we selected them as our audio representations. The extracted MFCC features are fed into a fully connected feed-forward network to get audio embeddings.

Text encoder. In our approach to processing text data, we begin by applying standard text processing procedures. These procedures include tokenizing the text data, converting all words to lowercase, removing English stopwords, and performing stemming and lemmatization on the words. We conducted experiments with various text extraction techniques, including one-hot encoding, bi-gram encoding, and the use of pre-trained text encoders such as BERT [29]. Our

results indicate that one-hot encoding provides the best performance compared to other encoding techniques. Hence, we use one-hot encoding as text representations. Similar to the audio encoder, the extracted one-hot encoding enters into a fully connected feed-forward network to get text embeddings.

Multimodal Co-attention Transformer Network. The visual, audio, and text representations are input into three symmetric multimodal co-attention sub-blocks. Each sub-block comprises a standard multi-head self-attention module and a proposed multimodal co-attention module. Layer normalization and a residual network are applied following each attention module, and a fully connected feed-forward network is also incorporated. The multi-head self-attention network independently identifies salient features from each modality. In contrast, the multimodal co-attention module learns the significant features of the interactions between the other two modalities, guided by the guiding modality. The residual network serves to stabilize the network and, more importantly, combines the guiding modality's representation with the joined representations, preserving information from all three modalities.

In the visual sub-block, the extracted visual embeddings from the visual encoder are first fed into a multi-head self-attention module to get the intermediate representation, I_v, containing important visual information. Similarly, the intermediate representations of audio and text, I_a and I_t, are obtained by feeding the extracted audio and text embeddings into the multi-head self-attention modules in their corresponding sub-blocks.

In the visual sub-block, the extracted visual embeddings from the visual encoder are input into a multi-head self-attention module to obtain an intermediate representation, I_v, containing salient visual information. Similarly, the intermediate representations for audio and text, I_a and I_t, are derived by inputting the extracted audio and text embeddings into their respective sub-

block multi-head self-attention modules. Specifically, I_v, I_a and I_t are calculated as follows:

$$\begin{aligned} I_v &= FC(Norm(MultiHead(z_v, z_v, z_v) + z_v)) \\ I_a &= FC(Norm(MultiHead(z_a, z_a, z_a) + z_a)) \\ I_t &= FC(Norm(MultiHead(z_t, z_t, z_t) + z_t)) \end{aligned} \tag{3.6}$$

where z_v z_a and z_t denote the output features from the visual encoder, the audio encoder and the text encoder respectively.

Next, in the visual sub-block, I_a and I_t are stacked to form a joined representation, denoted as $[I_a, I_t]$ with a dimension of $[2, d]$. Concurrently, the dimension of I_v is expended from 1-dimensional to 2-dimensional (e.g. from d to $[1, d]$). Subsequently, The multi-head co-attention module is then fed with I_v as the query and $[I_a, I_t]$ as both key and value. The resulting dot product between the query and key represents the similarities between audio, text, and visual embeddings with dimensions of $[1, 2]$, indicating the relative salience of audio and text embeddings with respect to visual embeddings. These similarity scores are then multiplied by $[I_a, I_t]$ to obtain a joined representation guided by the guiding modality, I_v. Similarly, joined representations of $[I_a, I_v]$ guided by I_t and $[I_v, I_t$ guided by I_a can be also obtained. Together with a residual network, layer normalization, and a fully-connected feed-forward network, joined representations for each modality (F_v, F_a, F_t) are calculated as follows:

$$
\begin{aligned}
F_v &= FC(Norm(MultiHead(I_v, [I_a, I_t], [I_a, I_t]) + I_v)) \\
F_a &= FC(Norm(MultiHead(I_a, [I_v, I_t], [I_v, I_t]) + I_a)) \\
F_t &= FC(Norm(MultiHead(I_t, [I_v, I_a], [I_v, I_a]) + I_t))
\end{aligned} \tag{3.7}
$$

Predictor. We apply a fully-connected feed-forward along with a sigmoid function to the output from the multimodal Co-attention network at the last step to make personality predictions, computed as follows:

$$
Personality = Sigmoid(FC(concat(F_v, F_a, F_t))) \tag{3.8}
$$

3.5 Experiments

3.5.1 Dataset

In order to assess the efficacy of our approach, we conducted experiments on a large-scale dataset: First Impressions [67]. The First Impressions dataset is a widely used benchmark in the field of apparent personality analysis and was employed in the ECCV 2016 personality trait recognition competition. It consists of 10,000 labeled video clips extracted from over 3,000 YouTube videos, with 6,000 designated for training and 2,000 for validation and testing. The dataset provides tri-modal information in the form of audio, visual, and text modalities. The average length of each video is 15 seconds and the majority have a resolution of $[1280, 720]$. Each video features a single individual speaking English in front of a camera. Ground truth

annotations for the Big-Five personality traits - extraversion, agreeableness, conscientiousness, neuroticism, and openness - are provided as fractional scores ranging from 0 to 1. The ECCV competition organizers obtained the annotations via Amazon Mechanical Turk. The summary statistics are shown in Table 3.1.

Table 3.1: Summary Statistics for First Impressions

Variables	Count	Mean	Standard Deviation	Min	Max
Agreeableness	10,000	0.55	0.13	0	1
Neuroticism	10,000	0.48	0.15	0	1
Extraversion	10,000	0.48	0.15	0	1
Openness	10,000	0.57	0.15	0	1
Conscientiousness	10,000	0.52	0.15	0	1
Video duration	10,000	15.28	0.47	2.04	15.45
Frame count	10,000	430.02	55.12	49.00	459.00

3.5.2 Implementation Details

The training of the model was conducted on an NVIDIA GeForce 3090Ti 24GB GPU for a total of 100 epochs. For the task of personality prediction, which is a regression problem with personality traits represented as continuous variables ranging from 0 to 1, the mean absolute loss was employed as the loss function during training, and Adam was utilized as the optimizer. For the task of emotion prediction, which is a binary classification problem, binary cross entropy was employed as the loss function and Adam was utilized as the optimizer. To prevent overfitting and improve model generalization, two early stopping mechanisms were implemented: training was halted if validation loss increased for five consecutive epochs or if validation loss failed to decrease for ten consecutive epochs. Additionally, a simple data augmentation strategy was employed during the training phase to further enhance the generalizability of the model. Specifically,

each image in the training set had a 50% chance of undergoing horizontal flipping.

The learning rate is set to $5e$-7. We perform a grid search method to select the best hyperparameter set. Specifically, the learning rate, batch size, number of heads, patch size, patch output dimension and the size of latent dimension were searched within the ranges of [$1e$-7, $5e$-7, $1e$-6, $5e$-6, $1e$-5, $5e$-5], $[8, 16, 32, 64]$, $[8, 16, 32, 64]$, $[(2, 2), (3, 3), (4, 4), (5, 5), (6, 6)]$, $[2, 4, 6, 8, 16]$ and $[128, 256, 512]$ respectively. In each experiment, the five best-performing models were saved for evaluation on the holdout test set.

3.5.3 Baselines

We compare the proposed model with the following baselines:

- **VATT** [3] is the most recent state-of-the-art model for handling vision, audio and text modalities. It uses a standard Transformer network to combine vision, audio and text modalities. Outperforming this baseline demonstrates the effectiveness of the proposed multimodal Co-attention network.

- **NJU-LAMDA** [102] is the best-performing algorithm in ECCV competition. It consists of a deep bi-modal regression model that combines visual features extracted from a Descriptor Aggregation Network and audio features extracted from a linear regressor. The authors propose two versions of Descriptor Aggregation Networks, DAN and DAN^{+}.

- **evolgen** [82] proposes two bi-modal networks. The first model, named evolgen-3D-Conv, is composed of a 3D-CNN for visual feature extraction, a linear regressor for audio feature extraction, and a fully-connected feed-forward network for combining both features. The second model, named evolgen-LSTM, divides each video into several equal segments.

Within each segment, an audio and an image are sampled and input into a linear regressor and a CNN, respectively. A fully-connected feed-forward network is employed to learn a joint representation for each video segment. These joint representations are then sequentially input into an LSTM to make predictions.

- **DCC** [36] develops two symmetric residual convolutional neural networks to extract audio and visual features that are fused by a fully-connected feed-forward neural network.

- **Multi-modal prediction** (MP) [83]. In this approach, each modality is trained independently. The visual encoder is fine-tuned from a pre-trained ResNet-101 model, while the audio encoder is fine-tuned from a VGG-16 model. A linear regressor is employed as a text encoder. The final prediction is obtained by taking the average of the predictions made by the three encoders.

3.5.4 Evaluation Metrics

Since the task is a regression task, we use the mean accuracy as the evaluation metric. Since the ground truth values are continuous, the accuracy of a video with its corresponding personality trait is computed as one minus the absolute distance between the predicted value and the ground truth value. Hence, the mean accuracy for a specific personality trait is computed as,

$$MeanAccuracy_t = \frac{1}{N}\sum_{i=1}^{N}(1 - |y_{it} - \hat{y}_{it}|) \tag{3.9}$$

where y_{it} is the ground truth value for the ith video sample and tth personality trait, and $\hat{y}_{it}$ is the predicted value for the same video sample and trait. N is the total number of predicted videos.

The mean accuracy across five personality traits is calculated in Equation 3.10,

$$MeanAccuracy = \frac{1}{5}\sum_{t=1}^{5}(MeanAccuracy_t) \tag{3.10}$$

3.5.5 Efficiency Metric

Efficiency is a critical consideration when training a deep learning algorithm. It is influenced by two primary factors: the total number of parameters and the total training time. These factors collectively determine the efficiency of a model. We propose an equation to quantify efficiency, which essentially measures the number of parameters that can be trained within a unit time. Mathematically, if we denote the total number of parameters as P and the total training time as T, the efficiency E can be calculated as follows:

$$E = \frac{P}{T} \tag{3.11}$$

In this context, a larger value of E indicates a more efficient model. This metric provides a standardized measure to compare the efficiency of different models, thereby facilitating more informed decisions in the selection and design of deep learning algorithms.

3.6 Results

3.6.1 Performance

Table 4.2 presents a comprehensive overview of the performance of our proposed model when compared to state-of-the-art baselines in the context of personality trait prediction. This

analysis demonstrates the superior capabilities of our model across multiple dimensions.

Our model exhibits remarkable consistency, achieving an accuracy of over 90% across all five personality traits and mean accuracy metrics, even when operating on low-resolution images (64x64 pixels). This achievement sets our model apart from the competition, as none of the state-of-the-art baselines are capable of such across-the-board success.

One distinguishing feature of our model is its utilization of audio MFCC information as opposed to raw audio waveforms used by the VATT model. Although both models share the same encoder networks, our approach proves more effective in capturing audio features. Furthermore, we employ a Multimodal Coattention Transformer network, which outperforms the standard Transformer network used by VATT in all evaluated metrics. Remarkably, our model achieves these results with approximately 3 million fewer parameters and faster convergence compared to VATT.

When compared to the NJU and DCC models, which utilize every sampled image from a video for training, our model stands out as the superior choice. This suggests our model's ability to effectively capture sequential and temporal information in videos, outperforming the alternatives. Notably, our approach reduces training time by operating at the video level rather than the sampled image level.

In contrast to the Evolegen-3D-Conv and Evolegen-LSTM models, which employ 3D convolutional neural networks and LSTMs, respectively, to capture temporal information in videos, our proposed model exhibits superior performance. This demonstrates our model's effectiveness in capturing temporal information. Interestingly, the LSTM architecture significantly outperforms 3D-Conv in predicting personality traits, highlighting the strength of LSTM in this context.

In comparison to the MP model, which relies on two pre-trained networks (ResNet and

Table 3.2: Performance comparison of our model with baselines for Big-Five personality predictions.

Method	Mean Accuracy	Extraversion	Agreeableness	Neuroticism	Conscientiousness	Openness
VATT [3]	0.889	0.891	0.887	0.886	0.890	0.891
NJU-DAN [102]	0.869	0.865	0.880	0.865	0.863	0.873
NJU-DAN^+ [102]	0.871	0.867	0.881	0.866	0.866	0.875
evolgen-LSTM [82]	0.891	0.891	0.897	0.889	0.884	0.894
evolgen-3D-Conv [82]	0.769	0.771	0.759	0.774	0.767	0.775
DCC [36]	0.885	0.881	0.892	0.883	0.879	0.888
MP [83]	0.878	0.880	0.883	0.876	0.873	0.879
Ours	**0.903**	**0.902**	**0.906**	**0.900**	**0.904**	**0.903**

VGG) to extract visual and audio features, our model is trained from scratch and still outperforms MP. This finding suggests that in this particular context, pre-trained models may not be as beneficial for prediction tasks. Moreover, the independent training of each modality in MP necessitates additional effort to integrate and interpret the results, further emphasizing the advantages of our approach.

In summary, our proposed model excels across various dimensions compared to state-of-the-art baselines, including accuracy and the ability to capture multimodal information effectively. These results underscore the strength and practicality of our model in the domain of personality trait prediction from multimodal data.

3.6.2 Efficiency

In the rapidly evolving landscape of deep learning, there is a noticeable trend towards the development of increasingly larger and more complex models. As a result, it becomes imperative to closely examine the efficiency of these models, considering not only their overall training time but also the number of parameters involved. To quantify this efficiency, we introduce a novel metric that measures the number of parameters that can be effectively trained per unit of time.

Table 3.3: Cost Analysis

Method	Number of Parameters	Average Training Time Per Epoch (min)	Total Training Time (min)	Efficiency
VATT [3]	849,099,017	3:16	163:20	519,856.41
NJU-DAN [102]	17,080,138	108:07	10807:23	1,508.41
NJU-DAN^+ [102]	19,177,290	93:01	9301:56	2,061.65
evolgen-LSTM [82]	136709	0:07	11:43	11,667.91
evolgen-3D-Conv [82]	105,538	0:11	15:35	6,772.49
DCC [36]	7,371,717	131	1572:16	4,688.59
MP [83]	191,431,031	1:24	41:59	4,559,691.09
Ours	845,959,433	1:26	53:02	**15,954,466.37**

Note: In the training process of NJU-DAN, NJU-DAN^+ and DCC models, each sampled image is used as a training instance. With 100 images sampled per video, the number of training instances increases from 6,000 to 600,000, resulting in a significant increase in training time for these three models. Furthermore, in NJU-DAN, NJU-DAN^+ and MP models, each modality is trained independently. The average training time per epoch is calculated by summing the average training time per epoch for each modality. The total training time is obtained by summing the total training time across three modalities.

With this metric in hand, we embark on a thorough investigation of model efficiency, the results of which are thoughtfully presented in Table 3.3.

This table offers a comprehensive and insightful perspective on the efficiency of both baseline models and our proposed model. It serves as a valuable resource for assessing their performance in terms of training time and parameter count.

Remarkably, the proposed model showcases a significantly higher level of efficiency compared to all the baseline models, underscoring its superiority in this regard. This efficiency boost can primarily be attributed to the innovative co-attention structure we have introduced, which inherently demands fewer parameters, as well as the parallel processing capabilities of the Transformer architecture.

Specifically, when compared to VATT, which relies on a conventional Transformer model, our proposed model outperforms it significantly in terms of efficiency. This observation suggests

that our co-attention structure is more efficient than the standard Transformer architecture. Furthermore, our model also surpasses the best-performing baseline, evolgen-LSTM, in terms of efficiency. This result indicates that our co-attention structure outperforms LSTM not only in terms of accuracy but also in efficiency.

An interesting finding in our analysis is the substantial difference in efficiency between evolgen-LSTM and evolgen-3D-Conv. It's worth noting that evolgen-3D-Conv employs a 3D convolutional neural network to capture temporal information, while evolgen-LSTM utilizes an LSTM network. This contrast underscores the efficiency of LSTM in both performance and parameter efficiency in the context of video-based personality prediction.

Another noteworthy observation pertains to the MP model. While it boasts higher efficiency compared to other baselines, it does so at the cost of modality-specific training, necessitating an additional step to aggregate predictions from each modality. This added complexity may not be desirable in practice, making our proposed model an attractive alternative that combines high efficiency with a streamlined approach.

In conclusion, the results presented in Table 3.3 shed light on the efficiency of the proposed model and baseline models, offering valuable insights into their trade-offs between parameter count and training time. Along with the performance presented in Table 4.2, our proposed co-attention structure stands out as an efficient and effective choice for multimodal personality prediction, demonstrating its superiority over traditional Transformer architectures and other baseline models in both accuracy and efficiency.

3.7 Ablation Study

We conduct additional experiments to investigate if the proposed hierarchical positional encoding is effective at predicting personalities. The results, shown in Table 3.4, indicate the effectiveness of the proposed positional encoding. For conscientiousness, the proposed hierarchical positional encoding increases the prediction accuracy by 0.44%. For the other four labels, the proposed hierarchical positional encoding increases the prediction accuracy by around 0.1%. Since the personality labels are continuous between 0 and 1, the improvements are significant.

Table 3.4: Comparison of Positional Encoding vs No Positional Encoding

Method	Extraversion	Agreeableness	Neuroticism	Conscientiousness	Openness
With positional encoding	0.902	0.906	0.900	0.904	0.903
Without positional encoding	0.901	0.905	0.899	0.900	0.902

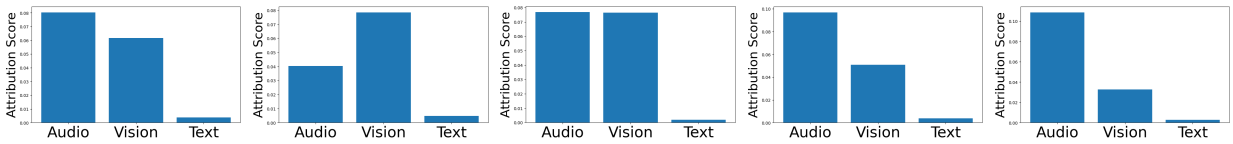
3.8 Interpretability Analysis

Furthermore, we conduct experiments to examine the contributions of features at different hierarchical levels to personality predictions. To achieve this, we used the Integrated Gradient method, as introduced by [88], to determine the attributions of each input, represented as $x \in \mathbb{R}^n$. This approach requires a finite number of gradient interpolation steps.

In particular, the Integrated Gradient (IG) for input x in the context of model F is computed as follows:

$$IG_i(x) := (x_i - x_i') \times \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m} \tag{3.12}$$

where m is the number of interpolation steps. $x' \in \mathbb{R}^n$ is the baseline, which is defined as an



(a) A (b) E (c) C (d) N (e) O

Figure 3.4: Average modality importance on personality prediction.

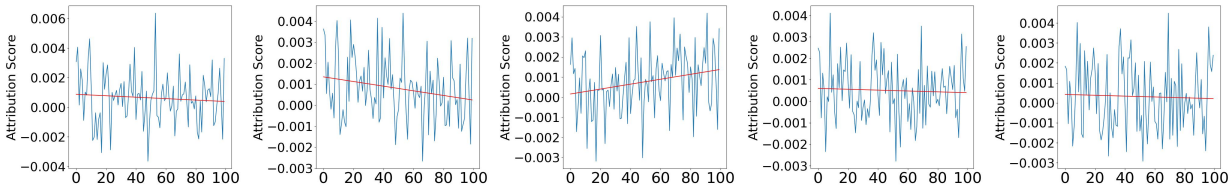
absence of input x. To be more specific, x' is the input from the input space that produces a neutral prediction.

This method enables the examination of the contributions of inputs at various levels by aggregating Integrated Gradient values. We use the whole training set to get contributions at three distinct levels:

- **Modality-level analysis** - This level investigates the average contribution of each modality (audio, vision, and text) to the predictions. The Integrated Gradients are aggregated for each modality. The results are presented in Figure 3.4.

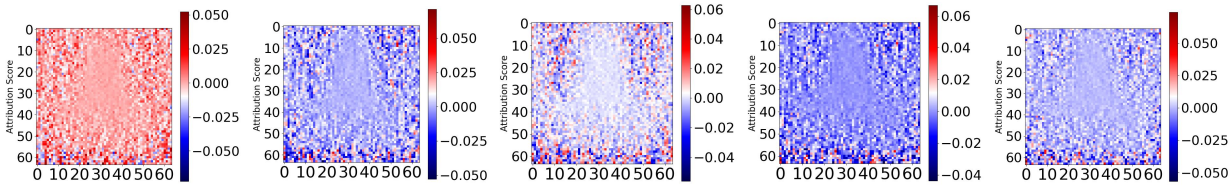
- **Vision-level analysis** - This level examines how different positions of pictures impact the predictions on average. The Integrated Gradients are aggregated for each vision by taking a summation of the value of each input across all images for each vision. The results are shown in Figure 3.5. To make the trend more salient, a linear regression is fitted for each personality trait.

- **Image-level analysis** - This level studies the average impact of different regions of an image with a resolution of 64×64 on the predictions. The Integrated Gradients are aggregated across images for each vision. The results are presented in Figure 3.6.



(a) A (b) E (c) C (d) N (e) O

Figure 3.5: Average contributions of images of different positions on personality prediction.



(a) A (b) E (c) C (d) N (e) O

Figure 3.6: Region contributions of images on personality predictions.

As depicted in Figure 3.4, our analysis reveals distinct patterns regarding the impact of different modalities on predicting the five personality traits. Here is a breakdown of our findings: General Information Modality: Surprisingly, the text modality, in a general sense, does not yield substantial information for predicting any of the five personality traits. Agreeableness: In the case of agreeableness, audio information slightly outperforms visual information, suggesting that sound-based cues carry a bit more weight in predicting this trait; Extraversion: Conversely, when it comes to extraversion, the model leans more heavily on visual information, indicating that visual cues are more influential in predicting this particular trait compared to audio cues; Conscientiousness: Notably, for conscientiousness, both audio and visual information contribute equally to the model's predictive power, signifying a balanced reliance on both modalities; Neuroticism and Openness: In the contexts of neuroticism and openness, the model primarily leans on audio information to inform its predictions, suggesting that auditory cues play a dominant role in shaping these personality trait predictions. In summary, our comprehensive analysis highlights a

nuanced interplay of modalities in personality trait prediction. While audio and visual information emerge as primary contributors, their relative importance varies depending on the specific trait, illustrating the intricacies of how different modalities inform the model's decision-making process.

Based on Figure 3.5, we can observe distinct patterns in how different personality traits are influenced by the temporal aspect of visual stimuli. Specifically: Agreeableness: The images located in the center of the visual field exhibit the most pronounced impact on predictions for agreeableness. However, this influence gradually wanes as time progresses; Extraversion: Extraversion, in contrast, shows a more conspicuous trend. The contributions of visual stimuli diminish consistently over time, with the images in the center retaining their informational value; Conscientiousness: Remarkably, the importance of images for predicting conscientiousness experiences a noteworthy increase as time advances; Neuroticism: In the case of neuroticism, the images presented at the outset have the most substantial influence on predictions. Nevertheless, this influence experiences a gradual reduction as time elapses; Openness: Similar to neuroticism, openness also exhibits a decline in the contributions of images as time passes, albeit the decrease is relatively minor. These findings underscore the dynamic relationship between personality traits and the temporal dynamics of visual stimuli, shedding light on the varying degrees of influence exerted by different portions of the visual field over time.

As illustrated in Figure 3.6, distinct regions of interest demonstrate comparable contributions across the predictions of the five personality traits. Notably, the regions situated in the middle of the visual field, where a person's face consistently resides, exhibit relatively less significance in shaping these predictions. In contrast, the surrounding regions, which encompass the background within a video frame, exert a more substantial influence on the perceived personality predictions. Remarkably, the regions located at the bottom portion of the frame, where most gestures and

actions typically occur, manifest the most pronounced contributions to the predictions. This observation underscores the pivotal role played by the movements of hands and arms in significantly impacting the predictions related to personality traits.

3.9 Decision Support Showcasing: MBA Admission Prediction

In this section, we explore the potential business insights generated by our video-based apparent personality detection model. We focus on assessing the impact of the generated Big-Five personality variables on MBA admission forecasting as a persuasion tool. To achieve this, we utilize MBA admission data from a prominent American university, where applicants are required to submit a video essay as part of the admission process. Figure A.1 in the appendix shows a screenshot of what the recording page looks like. The video essay question prompts vary. Applicants have 30 seconds to prepare their responses and up to 60 seconds to record their answers online. We aim to understand how the perceived personality of applicants, generated by our video personality predictive model, influences their admission decision for the full-time two-year program in 2022.

The two-year MBA program received a total of 895 applications, with acceptance rates at one-third. Summary statistics for each variable, namely the Big-Five personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism), as well as control variables like sex, age, international student status, average duration per job, number of jobs, test score, GPA and race, are presented in Table 3.5. Notably, applicants submitted either GMAT or GRE scores. To ensure consistency, GRE scores were converted to GMAT scores using the conversion table[1]. We additionally report the correlations among five personality variables, presented in Table 3.6.

[1]https://e-gmat.com/blogs/ gre-gmat-score-conversion/.

Table 3.5: Summary Statistics for the Case Study of MBA Admission

Variables	Count	Mean	Standard Deviation	Min	Max
Admission decision (Dependent Variable)	981	0.31	0.46	0	1
Agreeableness	981	0.52	0.05	0.41	0.69
Neuroticism	981	0.45	0.04	0.33	0.59
Extraversion	981	0.45	0.06	0.28	0.63
Openness	981	0.55	0.04	0.42	0.68
Conscientiousness	981	0.49	0.06	0.33	0.71
Sex	981	0.32	0.47	0	1
Age	981	28.14	3.98	20	45
If_native_English_speaker	981	0.30	0.46	0	1
If_international_student	981	0.29	0.45	0	1
Gpa_below_80	981	0.46	0.50	0	1
Gpa_above_80	981	0.51	0.50	0	1
No_gpa	981	0.03	0.17	0	1
No_gmat_score	981	0.32	0.47	0	1
Gmat_below_600	981	0.11	0.31	0	1
Gmat_between_600_and_680	981	0.17	0.37	0	1
Gmat_above_680	981	0.36	0.48	0	1
Number_of_jobs	981	2.35	1.24	0	5
Average_duration_per_job	981	1236.09	865.57	60	6787

Table 3.6: Personality Trait Correlations

Variables	Conscientiousness	Openness	Extraversion	Agreeableness	Neuroticism
Conscientiousness	1	0.58	0.67	0.69	-0.67
Openness	0.58	1	0.73	0.58	-0.78
Extraversion	0.67	0.73	1	0.70	-0.73
Agreeableness	0.69	0.58	0.70	1	-0.69
Neuroticism	-0.67	-0.78	-0.73	-0.69	1

The table shows the predicted personality traits are highly correlated with each other. To prevent the multicollinearity, we estimate the following equation:

$$ad_i = \alpha_0 + \alpha_1 personality_{i,j} + \vec{\alpha} controls_i + \epsilon_i \tag{3.13}$$

where ad_i represents whether applicant i is admitted to the MBA program. The explanatory variable, $personality_{i,j}$, denotes the perceived personality trait j of the video essay submitted by applicant i. We also include the control variables in the regression analysis. Since the dependent variable is a binary variable, we use logit to estimate the coefficients. As a robustness check, we

also report estimates using probit estimation.

Table 3.7: Estimation Results for MBA Admission

	Dependent Variable: Admission Decision											
Variables	Logit Estimation						Probit Estimation					
openness		1.75 (2.00)						1.09 (1.16)				
conscientiousness			1.98 (1.39)						1.17 (0.79)			
extraversion				2.71* (1.40)						1.69** (0.80)		
agreeableness					3.60** (1.72)						2.17** (0.99)	
neuroticism						-3.11 (1.98)						-1.91* (1.13)
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.202	0.203	0.204	0.205	0.206	0.204	0.203	0.203	0.205	0.206	0.207	0.205
Observations	981	981	981	981	981	981	981	981	981	981	981	981

Note: $*p < 0.1$, $**p < 0.05$, $***p < 0.01$. Our exploratory analysis has revealed that certain personality variables exhibit correlations with one another. To prevent multicollinearity, we elect to analyze each personality variable independently. The second and the eighth columns are the results without adding personality variables as dependent variables.

In Table 3.7, we present the logit estimation and probit estimation results, providing a comprehensive analysis of how personality variables influence the likelihood of admission to an MBA program. To better understand the impact of these variables, we initially present results that exclude personality variables as dependent variables in the second and eighth columns. However, we subsequently introduce these personality variables to assess their effects on predictive power.

First, by examining the R^2 values, we observe that incorporating personality variables enhances the models' predictive capabilities compared to those using only control variables. This suggests that personality traits indeed contribute significantly to explaining the variance in admission outcomes.

In the logit estimation, specific personality traits such as extraversion and agreeableness

emerge as noteworthy factors. The coefficient of extraversion is positively significant at the 90% confidence level, indicating that candidates perceived as more extroverted have a higher likelihood of admission. Similarly, the coefficient of agreeableness is positively significant at the 95% confidence level, further suggesting that agreeable individuals also enjoy a higher probability of being admitted. On the other hand, personality traits like openness, conscientiousness, and neuroticism do not demonstrate significant associations with admission outcomes in the logit estimation.

When we employ probit estimation, the significance levels of some coefficients change. Specifically, the coefficient of extraversion becomes significant at the 95% confidence level, strengthening the case for its positive association with admission. It is worth noting, however, that the coefficient of neuroticism approaches significance at the 89% confidence level, which becomes significant at the 90% level when using probit estimation. This suggests that being perceived as neurotic may indeed have some influence on admission outcomes. Meanwhile, the significance levels of openness, conscientiousness, and agreeableness remain consistent with the logit estimation results.

In summary, our results demonstrate robustness across different estimation methods. We can confidently conclude that being perceived as extroverted and agreeable is positively associated with a higher likelihood of admission. Notably, candidates perceived as agreeable seem to have an even stronger advantage, as indicated by the higher magnitude of their coefficient compared to extraversion and conscientiousness. However, the coefficients of openness and neuroticism are not significantly different from zero, suggesting that applicants' perception as open or neurotic does not significantly affect admission decisions. These findings provide valuable insights into the role of personality in the MBA admission process and have implications for both applicants

and admission committees, demonstrating the usefulness of the proposed model.

Table 3.8: Factor Analysis

Test	Conscientiousness	Openness	Extraversion	Agreeableness	Neuroticism
VIF	2.3362	3.1150	3.2084	2.5708	3.5437
KMO	0.90	0.82	0.87	0.86	0.85

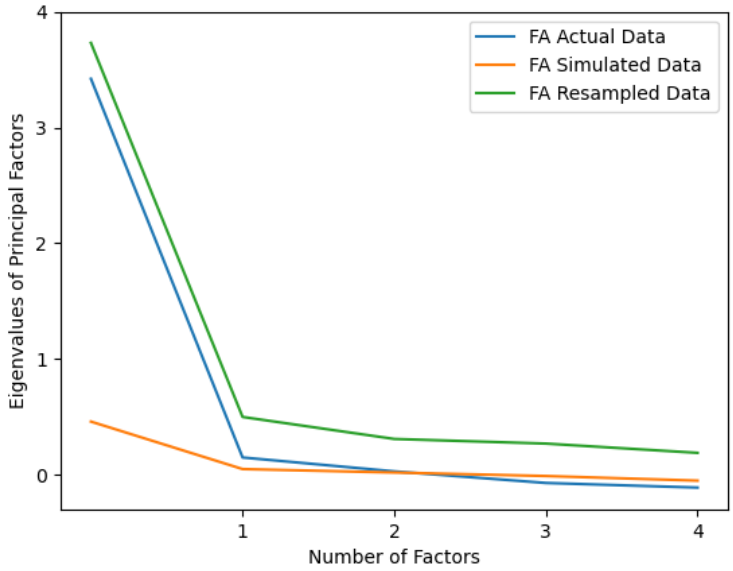


Figure 3.7: Parallel Analysis Scree Plots.

In addition, we conduct factor analysis to demonstrate that personality traits effectively capture latent factors capable of predicting MBA admission outcomes. To achieve this, we initially compute the Variable Inflation Factor (VIF) for each personality variable, as presented in the first row of Table 3.8. Based on the VIF values, we identify that variables such as Openness, Extraversion, Agreeableness, and Neuroticism warrant attention. Subsequently, we assess the factorability of the personality variables within our dataset using the Kaiser-Meyer-Olkin (KMO) Test, as indicated in the second row of Table 3.8. With an overall MSA value of $0.86 > 0.5$, we

proceed with factor analysis. To determine an appropriate number of factors, we conduct parallel analysis and generate a scree plot (Figure 3.7). The results suggest that one factor may be suitable. Employing a principal factor solution, we find that a single factor suffices. Consequently, we utilize this extracted factor to estimate the following equation:

$$ad_i = \alpha_0 + \alpha_1 factor_i + \vec{\alpha} controls_i + \epsilon_i \tag{3.14}$$

where ad_i represents whether applicant i is admitted to the MBA program. The explanatory variable, $factor_i$, denotes the extracted factor of the video essay submitted by applicant i. We also include the control variables in the regression analysis. The results are shown in Table 3.9.

Table 3.9: Estimation Results for MBA Admission Using Extracted Factor

Extracted Factor	Controls	R^2	Observation
0.16**	✓	0.206	981

Note: $**p < 0.05$.

The results reveal that the extracted factor achieves statistical significance at the 95% confidence level. This finding implies that the collective influence of the five personality traits produces a meaningful signal that can be effectively utilized to predict MBA admission outcomes. Additionally, the results suggest that the proposed model has the capability to generate variables that are valuable for further analysis and subsequent applications. This underscores the potential utility of the model in enhancing predictive accuracy and facilitating various downstream uses.

3.10 Discussion and Conclusion

In this research paper, we present a groundbreaking approach known as the Multimodal Co-attention Transformer network, designed to predict individuals' Big-Five personality traits by harnessing the rich information contained in audio, visual, and textual sources. Our innovative model surpasses the performance of existing methods, particularly when evaluated on the First Impressions dataset. This achievement underscores the remarkable effectiveness and efficiency of our Multimodal Co-attention mechanism in capturing intricate interactions among these different modalities. One of the notable advantages of our approach is its ability to extract valuable insights even from low-resolution image frames. This capability not only enhances the accuracy of personality trait prediction but also leads to substantial reductions in computational costs. This efficiency is pivotal in real-world applications where computational resources may be limited. Furthermore, our study introduces a novel hierarchical positional encoding technique, which serves to further bolster prediction accuracy. By incorporating additional relevant information into the model's architecture, we enable it to make more precise and contextually informed predictions, enhancing its overall performance. In addition to its predictive capabilities, our research also delves into the interpretability of the Multimodal Co-attention Transformer network. We conduct thorough analyses that shed light on how the model arrives at its decisions. This interpretability is of significant value not only to researchers seeking a deeper understanding of the model's inner workings but also to practitioners who wish to utilize it effectively in various applications. To showcase the practical utility of video-based personality detection using our multimodal approach and advanced attention mechanisms, we present a compelling case study. This real-world example demonstrates the model's accuracy in predicting Big-Five personality

traits through the integration of diverse data sources, offering practical implications for a wide range of applications. In summary, our research paper introduces a pioneering Multimodal Co-attention Transformer network that significantly advances the field of personality trait prediction by leveraging audio, visual, and textual information. With superior performance, reduced computational costs, enhanced accuracy, and valuable insights into interpretability, our model has the potential to revolutionize personality assessment across numerous domains.

Managerial Implications. Our comprehensive study yields valuable insights into actionable strategies that can significantly impact how influencers, online presenters, managers, and MBA program applicants are perceived by their respective audiences and evaluators. For influencers and online presenters seeking to project positive personalities characterized by attributes like agreeableness, openness, or extroversion, our findings suggest the importance of incorporating specific elements into their presentations. Firstly, they should focus on enhancing their body language, especially by increasing hand movements. These dynamic gestures not only capture the audience's attention but also convey a sense of enthusiasm and engagement. Furthermore, the choice of background plays a crucial role in shaping the perception of warmth and friendliness. Using warm and inviting backgrounds can create a more favorable impression, making viewers more likely to associate these influencers with positive traits. In addition to body language and backgrounds, the influence of voice should not be underestimated. The tone, cadence, and overall delivery of speech can significantly impact how individuals are perceived. A warm and expressive voice can reinforce the image of an agreeable and extroverted personality. For managers or platforms responsible for evaluating individuals online, our study underscores the importance of paying particular attention to specific factors. Hand movements, as previously mentioned, can reveal a lot about an individual's engagement and enthusiasm. The choice of background

should also be taken into consideration as it can influence perceptions of professionalism and approachability. Additionally, assessing the quality of an individual's voice and its congruence with the desired personality traits is essential when making evaluations. Furthermore, our case study indicates that individuals perceived as agreeable and extroverted have a higher likelihood of being admitted to MBA programs. As such, applicants aiming for success in their MBA interviews should leverage these insights to their advantage. By carefully selecting their backgrounds, modulating their voices, and incorporating appropriate hand movements, applicants can enhance their chances of projecting the desired traits and making a favorable impression on the interview panel. In conclusion, our research provides actionable strategies for various scenarios, emphasizing the significance of body language, background choices, and voice modulation in shaping perceptions. Whether aiming to build a positive online presence or seeking admission to competitive programs, individuals and organizations can benefit from implementing these insights to achieve their goals.

Limitations. While our study presents promising results, it is not without its limitations. Firstly, our proposed model does not explicitly consider temporal information, such as that captured by recurrent neural networks (RNNs), which have been shown to be crucial in video classification tasks. Instead, we utilize a Transformer network, which has demonstrated the ability to implicitly capture temporal information [97]. Future work will explore the potential benefits of integrating RNN-based networks into our model structure. Secondly, our model was trained on a single large-scale real-world dataset. The inclusion of additional datasets could enhance the generalizability of our model. Thirdly, our focus was on the MBA admission task. Investigating other video-based personality applications could further validate the utility of our proposed model. Despite these limitations, we believe our model has significant implications for research at the intersection of data science and social science. It enables researchers and practitioners to conduct predictive

analyses using video-based personality assessments, potentially leading to more insightful business decisions.

Chapter 4: Network-enhanced Multimodal Co-attention Learning for Short-Form Video Popularity Prediction

4.1 Introduction

The contemporary era has witnessed a remarkable worldwide escalation in the prevalence of short-form videos, propelled by video-sharing platforms including, TikTok, Instagram, and YouTube. For instance, TikTok alone reports a staggering 1.5 billion monthly active users as of 2024[1]. These platforms function as conduits for content generation and enable users to disseminate aspects of their lives, thereby promoting global interaction and communication. Importantly, the endeavor to predict video popularity on these platforms has garnered increasing attention, given its practical ramifications for various business applications such as strategic planning [107], targeted advertising [33], personalized recommender systems [20], and information filtering [90]. For instance,

Nonetheless, the popularity of videos can be influenced by a multitude of factors, encompassing aspects such as video quality, the subject matter, and the social network, among others. There has been a surge in interest in the prediction of video popularity. The majority of the research in this area has centered on the utilization of deep learning models, particularly those employing multimodal representation learning techniques, to learn from unstructured data, such as audio,

[1] https://sproutsocial.com/insights/tiktok-stats/

Figure 4.1: An illustration of a TikTok post.

Note: a TikTok post typically contains video, music and popularity information.

text, video, and network (both social and video networks) [21, 64, 95, 99, 115]. These models have underscored the utility of features derived from various modalities, including audio, video, and networks. However, these models are not without their limitations. Firstly, these deep learning models are incapable of capturing the interrelationships among different modalities, a factor that multiple multimodal learning studies have deemed crucial in making multimodal predictions [44, 84, 85]. Secondly, while research has highlighted the value of network information [21, 90, 112], existing models either rely exclusively on network information to make video

popularity predictions [90, 112] or incorporate hand-crafted network features to represent social information into a multimodal learning framework [21]. The former approach fails to integrate other types of information, such as audio and video information, which are also pivotal in popularity prediction, while the latter is unable to capture the overall network structure. Moreover, one of the emerging features of short-form platforms is the music network, wherein a video can opt to play a song that others have already selected, as depicted in Figure 4.1. This feature enables videos to be disseminated through the music network, thereby influencing the popularity of the video. To the best of our knowledge, no prior work has exploited this type of information in video popularity prediction, which we attempt to exploit.

In light of the challenges delineated above, this study, drawing inspiration from Sun and Zhang [84], proposes VAN: a multimodal learning framework capable of discerning relationships among **v**isual, **a**udio and **n**etwork modalities. Contrary to the model by Sun and Zhang, which is limited to processing three modalities, our model is designed to accept features from an arbitrary number of modalities and learn interrelationships among them. Furthermore, the proposed co-attention structure allows us to examine the inter-relationships among three modalities and compute the contributions of each modality as interpretability analysis.

In this context, we exploit visual, audio, and network information to predict popularity. To model visual information, we adopt the visual encoder proposed by [84], which has demonstrated efficacy in extracting information from low-resolution videos. For audio features, we employ a fully-connected feed-forward network (FC) to learn audio embeddings. To model network features, we adopt the Graph Attention Network (GAT) to learn network embeddings. Furthermore, we devise a network pre-training strategy to enhance the efficiency of the learning process. To assess the effectiveness of our model, we compile a dataset from TikTok, comprising 88,279

videos. The results not only underscore the superiority of the proposed model over state-of-the-art baselines but also illustrate a reduction in computational costs. We additionally conduct a comprehensive ablation study to understand the model behavior. Lastly, we conduct interpretability analysis to examine the contributions of each modality to the model performance. Our contributions can be summarized as follows:

- Our research introduces a generalized multimodal learning framework capable of accommodating an arbitrary number of modalities. Our architecture significantly enhances both its generalizability and efficiency over other multimodal architectures.

- We integrate a GAT network designed to model the bipartite video-music network, specifically focusing on capturing the network information associated with videos. Our GAT architecture effectively learns the relationships between music and video entities, enhancing the overall understanding of video content. Additionally, we introduce a pre-training strategy that significantly improves training efficiency, allowing our model to learn from large-scale data efficiently.

- Extensive experiments conducted on a real-world dataset unequivocally demonstrate the superiority and remarkable effectiveness of the proposed model.

- We exploit the unique characteristic of the co-attention structure to carry out an interpretability analysis. The findings from this analysis reveal that the audio and music network modalities contribute more significantly to the model's performance compared to the vision modality. This observation underscores the importance of these modalities in enhancing the effectiveness of the model and provides valuable insights for future research in multimodal learning.

4.2 Related Work

4.2.1 Online Video Popularity Prediction

The prediction of short-form video popularity has recently gained significant interest due to its substantial potential for business applications. This interest has been particularly pronounced in the context of the widespread application of neural networks and deep learning techniques in the fields of image processing, graph networks, and recommender systems. Consequently, numerous methods that leverage deep learning techniques have been proposed to predict the popularity of videos. These methods underscore the growing importance of this research area and its relevance to contemporary business practices. The first stream of research exclusively employs visual cues for the prediction of video popularity. Trzcinski et al. [95] suggest a Long Short-Term Memory (LSTM) network for the prediction of popularity in Facebook videos. In a similar vein, Wang et al. [99] present a pyramidal skeleton graph convolutional network for the prediction of popularity in dance challenge videos, utilizing visual cues that encompass skeletal, holistic appearance, facial, and scenic elements.

Furthermore, a significant portion of research has embraced multimodal learning for the prediction of video popularity. Chen et al. [21] devise a transductive multimodal learning model that predicts short video popularity by harnessing information from multiple modalities, including text, visuals, acoustics, and social data. Zhu et al. [115] put forward a variational encoder-decoder approach that amalgamates information from multiple modalities, such as text, audio, vision, and social data, for the prediction of video popularity. Nevertheless, these two methodologies exclusively take into account metrics such as follower count and post frequency as

social indicators. Unfortunately, these particular social cues fail to encapsulate the comprehensive network structure. In addition, Ou et al. [64] suggest a multimodal and temporal attention fusion network that integrates both visual and textual information.

Lastly, certain studies rely solely on graph networks for the prediction of short video popularity. Zhang et al. [112] implement a region-based graph neural network for video popularity prediction. Tang et al. [90] utilize a knowledge-based heterogeneous graph neural network and employ an attention-based LSTM to learn the temporal dependency of features for video popularity prediction.

Although these methodologies underscore the significance of various information sources, they either necessitate feature engineering for feature generation or solely rely on a single type or a subset of information for predictions. Unfortunately, both approaches can adversely impact the model's performance. In contrast, our proposed model addresses this issue by constructing a cohesive framework that incorporates visual, acoustic, and social structural aspects. Notably, our model avoids reliance on feature engineering, granting it greater flexibility. Specifically, we incorporate a graph neural network into our framework, enabling the model to learn the underlying social structure of the network. This integration significantly enhances the model's performance.

4.2.2 Multimodal Representation Learning

Multimodal representation learning is a deep learning technique that aims to create a joint representation from multiple modalities, such as images, text, audio, and video. This approach has garnered significant attention due to its potential for downstream applications. Early research

focused on straightforward combination methods to integrate information across modalities. For example, Kuhnke et al. [47] employed a fully connected layer to merge aural and visual embeddings for video emotion prediction. Poria et al. [68] concatenated acoustic, visual, and textual embeddings, using kernel learning to predict emotions and sentiments. Zhou & Shen et al. [114] conclude that a simple linear combination of modalities can achieve good results in multi-view clustering. However, this approach may result in information loss [87].

Recent work emphasizes the importance of learning interactions among modalities, with the Transformer architecture [97] being widely adopted as the backbone network for this purpose. Tang et al. [91] propose a vision-language pre-training architecture with a Transformer network, showing great performance in downstream applications, such as visual question answering, and video retrieval. Sun and Zhang [84] develop a multimodal architecture to learn the relationships among modalities and make personality predictions using a Tranformer-based co-attention network. Cheng et al. [22] develop a Tranformer-based cross-attention model to learn a joint representation between audio and vision and predict their synchronization. Sun and Zhang [85] propose a Transformer co-attention to learn a joint representation between audio and video at the segment level and use an LSTM to learn the temporal dependency of every segment.

Current multimodal learning architectures primarily center around acquiring a cohesive representation that integrates audio, video, and text, mirroring the way human cognition processes information. However, existing models have largely overlooked the influence of network information in multimodal learning. Our proposed model addresses this gap by incorporating network-related features into the architecture, thereby enhancing the overall multimodal representation learning process.

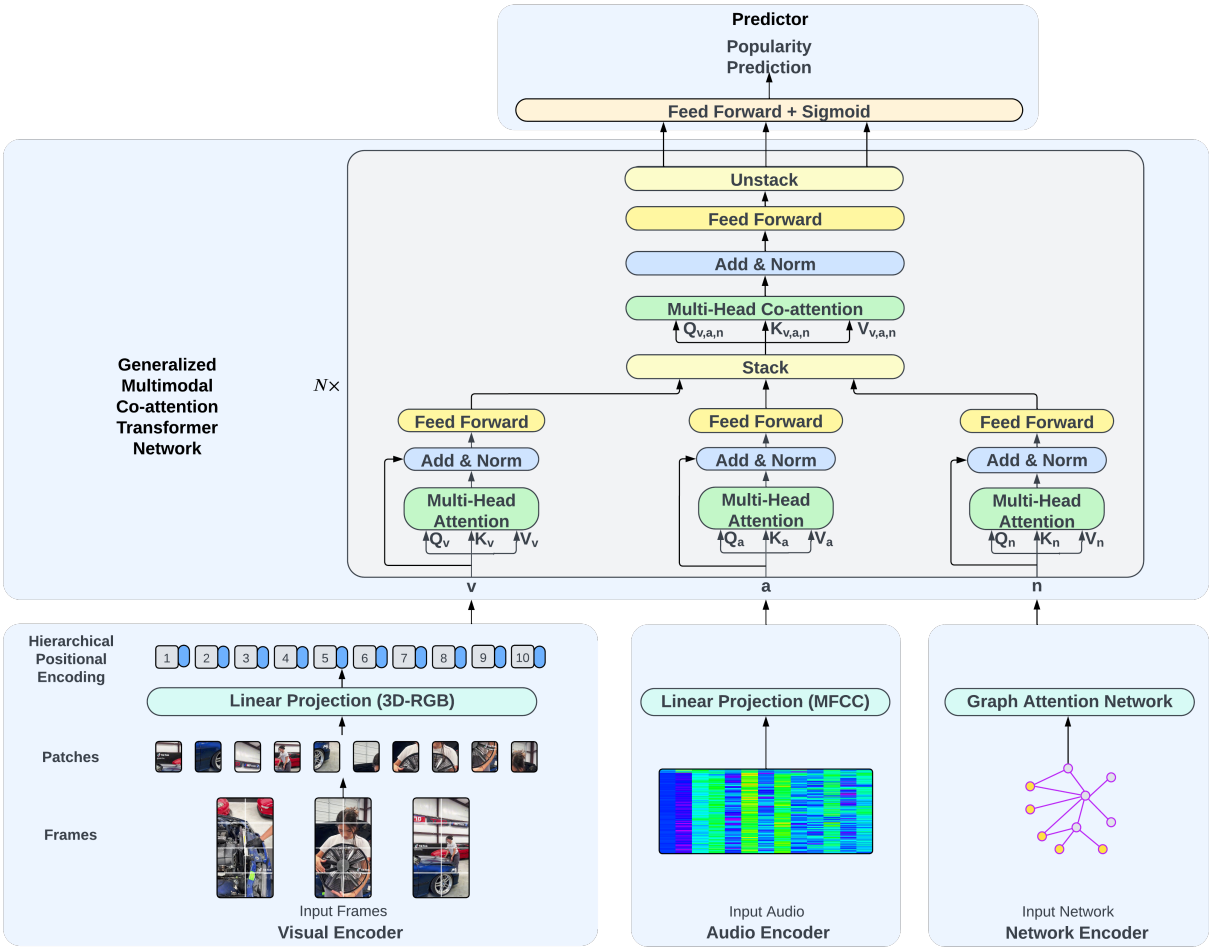


Figure 4.2: An overview of our proposed model: A Generalized Multimodal Co-attention Transformer.

4.3 Proposed Model

In this section, we present VAN: A Generalized Multimodal Co-attention Using **V**isual, **A**coustic and **N**etwork Cues, an efficient Co-attention Transformer model that accepts a list of embeddings from multiple modalities, as depicted in Figure 4.2.

4.3.1 Input Embeddings

The input embeddings of the VAN model are composed of three distinct types: visual embedding, audio embedding, and network embedding. As depicted in Figure 4.2, these three categories of embeddings are produced by three respective trainable encoders: the vision encoder, the audio encoder, and the network encoder. We now explain the details of these three types of embeddings.

Visual Embedding. We adopt the visual encoder proposed by Sun and Zhang [84], an extension of Vision Transformer [31]. This encoder is composed of two primary components: a visual encoder for encoding visual data, and a hierarchical positional encoding mechanism for the incorporation of two forms of spatiotemporal information, namely spatial information derived from patches and temporal information extracted from frames. To encode frames, we adhere to the standard procedure as outlined in [31, 39]. Each frame, originally sized at 576×1024 pixels, is resized to a dimension of 224×224 pixels. Subsequently, these resized pixels are segmented into a series of patches, each of size $P \times P$. A linear projection layer is then applied to the normalized pixel values within each patch. For a video comprising N frames, the input, with a shape of $N \times 3 \times 224 \times 224$ (representing time, channel, height, and width respectively), will yield $N \times \frac{224}{P} \times \frac{224}{P}$ patches. It should be noted that N and P are hyperparameters that necessitate tuning.

To encode the temporal information of frames within a video, as well as the spatial information of patches within a frame, we utilize the hierarchical positional encoding as proposed in [84]. This encoding is constructed upon the Transformer model [97]. In the original instantiation of the Transformer model, positional encoding is achieved through the application of sine and cosine

encoding functions, denoted as $PE_{pos,i}$:

$$PE_{pos,2i} = sin(pos/10000^{2i/d}) \tag{4.1}$$

$$PE_{pos,2i+1} = cos(pos/10000^{2i/d}) \tag{4.2}$$

where pos pos denotes position and i represents the dimension. These trigonometric functions can be employed to encode two distinct types of positions: the position of the frame, represented as f, and the position of the patch, represented as p. These are respectively encoded as $PE_{f,i}$ and $PE_{p,i}$. Consequently, we can define the hierarchical positional encoding for the temporal information of the frame f and the spatial information of the patch p as $PE_{p,f,i}$ calculated as

$$PE_{p,f,i} = PE_{f,i} + PE_{p,i} \tag{4.3}$$

Audio Embedding. In the process of obtaining audio embeddings, the initial step involves the extraction of the raw audio signals from each video, which are stored as .wav files. Subsequently, the audio waveform is derived from the aforementioned audio signal. The next phase of the process entails the computation of the 2-dimensional Mel-frequency cepstral coefficients (MFCCs) [27] from the audio waveform. In the final stage, the MFCCs are flattened into a 1-dimensional space, followed by the application of a linear projection to obtain the audio embeddings. This methodology aligns with recent studies that have determined the effectiveness of applying a linear projection to flattened MFCCs in capturing pertinent information [84, 101]. We experiment with two audio feature extraction methods: MFCCs and log filter banks [75], and two sampling rates: the original 44.1kHz and the standard 16kHz.

Network Embedding. To derive the network embedding, we leverage the available information about the songs utilized in each video. In particular, we first construct a bipartite music-video network graph, denoted as $\mathcal{G} = \{(s, v)|s \in \mathcal{S}, v \in \mathcal{V}\}$, where $\mathcal{S}$ and $\mathcal{V}$ denote the song and video sets. An edge $e_{s,v} = 1$ indicates video v utilizes song s as the background music; $e_{s,v} = 0$ otherwise. Subsequently, this graph is inputted into a Graph Attention Network (GAT) as per the methodology proposed by Velickovic et al. [98]. This approach enables us to effectively capture and represent the complex relationships within the network.

We now describe how the information is aggregated through the GAT network. The input to GAT is two sets of node features: raw song features and raw video features, denoted as $\mathbf{h}^{\mathcal{S}} = \{\vec{h}_1^{\mathcal{S}}, \vec{h}_2^{\mathcal{S}}, ..., \vec{h}_N^{\mathcal{S}}\}$ and $\mathbf{h}^{\mathcal{V}} = \{\vec{h}_1^{\mathcal{V}}, \vec{h}_2^{\mathcal{V}}, ..., \vec{h}_M^{\mathcal{V}}\}$ respectively, where N is the number of songs and M is the number of videos. Each has a dimensionality of d_s and d_v respectively. In the initial step, two separate linear transformations are applied to two types of nodes. These transformations are parameterized by two weight matrices, denoted as $\mathbf{W}^{\mathcal{S}}$ and $\mathbf{W}^{\mathcal{V}}$. Following this, an attention mechanism is employed to learn the importance of video v to song s, parameterized by a weight vector $\vec{a}$. Lastly, we normalize the importance scores via a softmax function. Together, we have:

$$\alpha_{s,v} = \frac{exp(\sigma(\vec{a}^T[\mathbf{W}^{\mathcal{S}}\vec{h}_s^{\mathcal{S}}||\mathbf{W}^{\mathcal{V}}\vec{h}_v^{\mathcal{V}}]))}{\sum_{k\in\mathcal{N}_s} exp(\sigma(\vec{a}^T[\mathbf{W}^{\mathcal{S}}\vec{h}_s^{\mathcal{S}}||\mathbf{W}^{\mathcal{V}}\vec{h}_k^{\mathcal{V}}]))} \tag{4.4}$$

where $\sigma(\cdot)$ denotes the activation function, $||$ denotes the concatenate operation, $\mathcal{N}_s$ all the neighbors of song s and $\vec{a}$ is the node level attention vector. Note that $\alpha_{s,v}$ is an asymmetric weight. This implies that the contributions made by different entities to each other are not uniform, but rather depend on their features. This is due to the fact that these nodes have both different types and quantities of neighbors. Consequently, the embedding of a song s can be

computed by aggregating the projected features of its neighbors, along with their corresponding coefficients. This aggregation is achieved by employing the normalized attention weights and applying a nonlinear function, denoted as $\sigma(\cdot)$:

$$z_s = \sigma(\sum_{j \in \mathcal{N}_s} \alpha_{s,j} \mathbf{W}^{\mathcal{V}} \mathbf{h}_j^{\mathcal{V}}) \tag{4.5}$$

where z_s is the learned embedding of song s. Since each video is only connected with one song, the embedding of each video is equal to its corresponding song.

To further stabilize the learning process, we also adopt the multi-head attention mechanism proposed by Vaswani et al. [97]. Specifically, we learn K independent attention transformations which are concatenated to form the final embedding:

$$z_s = ||_{k=1}^{K} \sigma(\sum_{j \in \mathcal{N}_s} \alpha_{s,j}^{k} \mathbf{W}_k^{\mathcal{V}} \mathbf{h}_j^{\mathcal{V}}) \tag{4.6}$$

where K is the total number of heads, $||$ denotes the concatenate operation, $\alpha_{s,j}^{k}$ are the normalized attentions of the kth attention mechanism with the corresponding linear transformation weight matrix $W_k^{\mathcal{V}}$.

4.3.2 VAN: A Generalized Multimodal Co-attention Network

Transformer. The VAN network is built upon the standard Transformer architecture. We now briefly describe the Transformer model. In a standard Transformer, the key component is multi-head self-attention (MSA) that accepts query (Q), key (K) and value (V) as inputs and generates

an embedding, which is calculated as:

$$MSA(Q, K, V) = Concat(h_1, h_2, ..., h_n)W^O \tag{4.7}$$

where

$$h_i = softmax(\frac{(W_i^Q Q)(W_i^K K)^T}{\sqrt{d_{model}/n}})(W_i^V V) \tag{4.8}$$

where Wi^Q, Wi^K and $Wi^V \in \mathbb{R}^{d_{model}/n \times d_{model}/n}$ are trainable weights of each head i. $W^O \in \mathbb{R}^{d_{model} \times d_{model}}$ is the projection weight. d_{model} is the hidden dimension of the model. In addition, a fully connected feed-forward (FC) network is applied after the MSA layer. A residual connection and a normalization function are applied as well. Therefore, the Transformer model can be expressed as:

$$Transformer(z) = FC(Norm(MSA(z, z, z) + z)) \tag{4.9}$$

where z is the input of the model.

VAN. Drawing upon the concepts presented in [84], we have designed a generalized multimodal co-attention Transformer network. This network is capable of accepting embeddings from an arbitrary number of modalities and discerning the interrelationships among these modalities. This approach allows for a more holistic understanding of the relationships across multiple modalities. In the context of this study, the VAN network receives the visual, audio, and network embeddings, denoted as v, a and n respectively. These embeddings are initially processed by a standard Transformer structure, as described in [97]. Each modality is associated with its own distinct Transformer structure, which generates intermediate embeddings, denoted as $z_v = Transformer(v)$, $z_a = Transformer(a)$ and $z_n = Transformer(n)$ respectively. It is important to note that the

weights of each Transformer are not shared among the modalities. Then, z_v, z_a and z_n are stacked to form a joint representation of a two-dimensional matrix z_j with a dimensionality of $3 \times d_{model}$. Specifically, z_j can be expressed as follows:

$$z_j = [z_v; z_a; z_n] \tag{4.10}$$

Subsequently, the stacked representation z_j is fed into a multi-head co-attention network. This network employs the same computational approach as the MSA network. Additionally, an FC network, a residual connection, and a normalization function are incorporated into the process. Consequently, the output F_j from the multimodal co-attention network is calculated as follows:

$$F_j = FC(Norm(MSA(z_j, z_j, z_j) + z)) \tag{4.11}$$

where the product of $W^Q Q_{z_j} \times W^K K_{z_j}$ yields a 3×3 matrix, representing the interralationships among three modalities. F_j is a $3 \times d_{model}$ matrix, where each row corresponds to the embedding of a specific modality. An unstacking operation is performed to get embedding for three modalities, with each having a dimensionality of d_{model}.

In the final stage of our process, we concatenate three embeddings to yield a vector of dimensionality $3 \times d_{model}$. Subsequently, we employ an FC layer and a sigmoid activation function to generate a prediction for video popularity.

4.4 Experiments

4.4.1 Dataset

4.4.1.1 Data Collection

In order to carry out our experiment, we gather a dataset from the real-world platform, TikTok. Initially, we collect a list of the most frequently played songs on TikTok, sourced from Tokboard[2]. This list comprises 474 songs. After excluding songs that are not available in the United States, we are left with a total of 311 songs. We then retrieve lists of video IDs that utilize each song. From each list, we randomly select 500 videos. For every video, we download both the video itself and its associated metadata. This metadata includes the number of times the video was shared, the number of comments on the video, the number of times the video was played, the time the video was published, and so forth. After excluding videos or their metadata that are unavailable, we are left with a total of 88,279 videos.

4.4.2 Popularity Score

We adopt a widely adopted popularity score calculation method proposed by [21, 99]. Our popularity score is based on three related popularity measures: number of play counts, number of share counts and number of comment counts, which are denoted as n_{play}, n_{share} and $n_{comment}$ respectively. As the popularity of each video is accumulated over time, we normalize the popularity score with the time difference between the scraping time and the video v posted time (in days), denoted as t_v. We also apply a log transformation to handle the large variation and

[2]https://tokboard.com/

skewness of the original distribution, as suggested by [15]. Therefore, the popularity score p_v is defined as:

$$p_v = log(\frac{n_{play} + n_{share} + n_{comment} + 1}{t_v}) \tag{4.12}$$

The addition of 1 prevents the computation of a logarithm of zero. We further normalize each popularity score between 0 and 1 as follows:

$$p_{norm,v} = \frac{p_v - min(\vec{p})}{max(\vec{p}) - min(\vec{p})} \tag{4.13}$$

where $\vec{p}$ denotes all the popularity scores in the dataset.

4.4.3 Evaluation Metrics

Following the popularity prediction literature [21, 99], we employ Spearman's rank correlation as our metric for evaluation. Spearman's rank correlation is a non-parametric measure of rank correlation or statistical dependence between the rankings of two variables. It assesses how well the relationship between two variables can be described using a monotonic function. Spearman's rank correlation of 1 indicates a perfect positive correlation, meaning that the ranking order is identical to the ground truth. Conversely, Spearman's rank correlation of -1 signifies a perfect negative correlation, indicating that the ranking order is the exact opposite of the ground truth. This metric is particularly useful in our research as it allows us to quantify the degree to which our model's predictions align with the actual data.

Table 4.1: Hyperparameters of the proposed model.

Hyperparameter	Numerical Value
Number of epochs	100
Optimizer	Adam
Learning rate	5e-6
Batch size	256
Number of sampled frames	2
Patch size	32×32
Number of heads	4
Embedding size	1024

4.4.4 Implementation Details

For all the experiments, we use the Adam optimizer [45] with a learning rate equal to 5e-6. The model is trained on an NVIDIA GeForce 3090 24GB for 100 epochs. We use a batch size of 256 for each epoch training. We adopt mean square error (MSE) as the loss function. We implement an early stopping mechanism if the validation loss fails to decrease for 10 consecutive epochs. We adopt a 70-10-20 split, in which 70% of the data are used for training, 10% of the data are used for validation, and the rest of the data are used for testing. Furthermore, we apply grid search to find the optimal set of hyperparameters. Specifically, learning rate, batch size, the number of sampled frames, patch size, number of heads, and embedding size are searched within ranges of [1e-4, 5e-4, 1e-5, 5e-5, 1e-6, 5e-6], [18, 32,...,256], [1, 2,..., 15], [4×4, 8×8, 16×16, 32×32], [4, 8, 16, 32, 64, 128] and [256, 512, 1024, 2048] respectively. The hyperparameter settings are shown in Table 4.1. In each experiment, the five best-performing models were saved for evaluation on the test set.

4.4.5 Graph Attention Network Pre-training

The proposed model can be trained end-to-end. However, due to computational limitations, we adopt a link prediction pre-training strategy to make the training process more efficient, which is a commonly used pre-training strategy in graph network learning [50]. Specifically, the link prediction aims to predict the likelihood of a link existence between two nodes using GAT parameterized by Φ. For a pair of a song s and a video v, the probability of a link existence, $p(s, v)$ is thereby given by:

$$p(s, v) = GAT(s, v|\mathcal{G}, \mathbf{h}^{\mathcal{S}}, \mathbf{h}^{\mathcal{V}}, \Phi) \tag{4.14}$$

where $\mathbf{h}^{\mathcal{S}}$ and $\mathbf{h}^{\mathcal{V}}$ are corresponding feature sets for songs and videos respectively. The experimental setup involves a random allocation of 80% of the links for training purposes, 10% for validation, and the remaining links are utilized for testing. During the training phase, 60% of the links in the training set serve as training instances, while the remaining 40% are used as ground-truth labels. This distribution results in 47,670 training instances and 31,780 ground-truth labels. Furthermore, we incorporate an additional 23,835 negative examples, which constitute 50% of the positive links. The MSE loss is adopted as the loss function for this experiment. Upon completion of the training, we observe an MSE loss of 0.0148 in the testing set. After the pre-training phase, the GAT network is frozen. We only train an FC layer added on top of GAT.

4.4.6 Baselines

ViViT [9] represents the current state-of-the-art of video vision representation learning. It has demonstrated superior performance across a variety of downstream tasks. For this baseline, we utilize video data to train a ViViT model.

GAT [98] is a neural network architecture specifically designed to operate on graph-structured data1. It leverages masked self-attention layers to weigh the importance of each node's connections, making it particularly adept at handling large and complex graphs. We train a GAT network using our dataset in a semi-supervised manner.

Audio-MLP [60] is a multi-layered perceptron-based deep learning framework for audio classification. We train an Audio-Linear network using audio data.

BDAE [52] proposes a **B**i-modal **D**eep **A**uto**E**ncoder (BDAE) by leveraging a weighted sum fusion technique to fuse features from two modalities. To find the optimal combination of weights, a grid search is conducted. This approach has yielded state-of-the-art results for five video-based emotion tasks. We utilize our dataset to train a BDAE model.

BMT [41] proposes a **B**i-**M**odal **T**ransformer Encoder by employing a Transformer architecture for bimodal input. It has demonstrated exceptional performance on audio-visual caption tasks.

NJU-Lamda [101] suggests an ensemble approach to combine predictions from different models (modalities). Specifically, it advocates for training different modalities separately and averaging the predictions generated from these different modalities. This method has achieved state-of-the-art performance in video personality tasks. We adopt the ensemble method proposed in this work as a baseline.

MMVED [107] is the leading algorithm for video popularity prediction. It proposes a

Table 4.2: Performance comparison of our model with baselines.

Method	Vision	Audio	Network	Rank Correlation	Average Training Time Per Epoch (min)	Total Training Time (hour)	Number of Parameters
ViViT [9]	✓			0.4264	57.14	20.33	2,577,421
Audio-MLP [60]		✓		0.6289	33.73	13.43	275,187,713
GAT [98]			✓	0.6664	13.48	2.49	185,636,866
BDAE [52]	✓	✓		0.6835	37.62	62.70	54,987,279
BMT [41]	✓	✓		0.6927	37.24	15.51	128,847,885
NJU-Lamda [101]	✓	✓	✓	0.6045	68.54	37.98	463,402,000
MMVED [107]	✓	✓	✓	0.7198	30.67	16.63	141,876,237
SiMVC [94]	✓	✓	✓	0.7113	34.12	31.28	285,288,464
VATT [3]	✓	✓	✓	0.7147	39.31	12.67	323,063,821
Ours	✓	✓	✓	**0.7390**	35.54	11.93	302,633,997

Note: The NJU-Lamda model employs an ensemble technique wherein each modality is trained independently. The reported metrics, including the average training time per epoch, total training duration, and the number of parameters, are computed by taking a summation of the corresponding values from each modality.

stochastic multimodal encoder by leveraging a variational autoencoder architecture. When combining different modalities, it assumes probability independence among features from different modalities. This approach has achieved state-of-the-art results on video popularity prediction tasks.

SiMVC [94] suggests a linear combination of features from different modalities. This straightforward yet effective approach has demonstrated comparable or even superior performance on some tasks while maintaining superior computational efficiency.

VATT [3] is a state-of-the-art algorithm that leverages the Transformer structure to combine three different modalities: vision, audio, and text. In our work, we substitute the text modality with network information.

4.5 Results

Table 4.2 provides a comprehensive summary of the primary results of the proposed model, in comparison to those of the state-of-the-art baselines. In addition to this, we have undertaken a

thorough cost analysis to contrast the efficiency of the proposed model with that of the baselines. On the whole, the proposed model attains state-of-the-art results, surpassing all the baselines, while preserving a comparatively lower total training time than most of the baselines.

The model under consideration demonstrates superior performance over ViViT, Audio-MLP, and GAT, with improvements of 73.73%, 17.5%, and 10.9% respectively. This underscores the insufficiency of relying on a single modality for accurate video popularity prediction. Moreover, the model surpasses BDAE and BMT, two bi-modal networks, by margins of 8.1% and 6.7% respectively. Notably, GAT alone attains a rank correlation of 0.6664, which implies the significance of network information in predicting video popularity. The results collectively suggest that all three types of information are indispensable for accurate video popularity prediction, necessitating a robust multimodal network. Furthermore, the proposed model exhibits greater efficiency than the two bi-modal networks, particularly BDAE, as evidenced by its shorter training time.

The model proposed shares common encoder networks with four existing models: VATT, SiMVC, MMVED, and NJU-Lamda. When compared to these models, the proposed model demonstrates superior performance, outperforming the other four baselines by as much as 22.2%, while also requiring the least amount of training time. The proposed model surpasses NJU-Lamda by 22.2%, underscoring the critical role of fusion in multimodal learning for predicting video popularity. In the case of NJU-Lamda, each modality is trained independently, necessitating additional time for aggregating the results, which also incurs significant computational expense. The proposed model also outperforms MMVED, currently the best video prediction model, by 3%. This suggests that, in this context, a deterministic model is more effective than a stochastic model for predicting video popularity. Compared to SiMVC, the proposed model shows nearly 4% improvement while requiring significantly less training time. This indicates that the proposed

Table 4.3: Ablation study of the proposed model.

Audio Feature	Audio Sampling Rate	Number of Heads	Embedding Size	Frame Patch Size	Positional Encoding	Rank Correlation
MFCCs	16kHz	4	1024	32×32	✓	**0.7390**
MFCCs	44.1kHz	4	1024	32×32	✓	0.7285
Log Filter Banks	16kHz	4	1024	32×32	✓	0.6906
Log Filter Banks	44.1kHz	4	1024	32×32	✓	0.6817
MFCC	16 kHz	4	256	32×32	✓	0.7356
MFCC	16 kHz	4	512	32×32	✓	0.7362
MFCC	16 kHz	4	2048	32×32	✓	0.7326
MFCC	16 kHz	8	1024	32×32	✓	0.7381
MFCC	16 kHz	16	1024	32×32	✓	0.7379
MFCC	16 kHz	32	1024	32×32	✓	0.7372
MFCC	16 kHz	64	1024	32×32	✓	0.7368
MFCC	16 kHz	128	1024	32×32	✓	0.7364
MFCC	16 kHz	4	1024	4×4	✓	0.7325
MFCC	16 kHz	4	1024	8×8	✓	0.7303
MFCC	16 kHz	4	1024	16×16	✓	0.7342
MFCCs	16kHz	4	1024	32×32	✗	0.7314

model converges faster than SiMVC, saving approximately five hours of training time. Finally, the proposed model outperforms VATT by 3.4%. Although both models require a similar amount of training time, the proposed model has more than 20 million fewer parameters than VATT. In conclusion, the proposed model excels over these four baselines in learning the relationships among three modalities and making accurate predictions.

4.6 Ablation Study

In Table 4.3 and Figure 4.3, a comprehensive series of ablation studies are conducted to substantiate the effectiveness of the pivotal components incorporated in the proposed model. These components include audio feature extraction (row 2∼4), the dimensionality of the latent embedding (row 5∼7), the number of attention heads (row 8∼12), patch size (row 13∼15), positional encoding (row 16) and the quantity of sampled frames (Figure 4.3). The first row is the optimal hyperparameter set. The results serve to validate the integral role these elements play in the overall performance of the model.

Audio Feature Extraction. The two methodologies that have gained the most popularity are Mel Frequency Cepstral Coefficients (MFCCs) [27] and logarithmic filter banks (logfbanks) [75]. These methodologies have been utilized extensively in a variety of audio classification tasks [84, 101]. Furthermore, two distinct sampling rates are frequently employed for the extraction of audio features: the original rate of 44.1kHz and the standard rate of 16kHz. The results for the two aforementioned audio feature extraction methodologies and the two sampling rates are presented in rows 2 through 4. It is observed that, with the same sampling rate of 16kHz, MFCCs yield substantial performance improvements (approximately 7%) in comparison to logfbanks. With regard to the sampling rate, the standard rate of 16kHz outperforms the original rate of 44.1kHz for both MFCC and logfbank features. Specifically, for MFCCs, the standard 16kHz rate is 1.5% superior to the original 44.1kHz rate. For logfbanks, the standard 16kHz rate is 1.3% superior to the original 44.1kHz rate.

Embedding Size. A crucial aspect of the model's design pertains to the dimensions of the latent embedding output derived from each encoder. We scrutinize the performance of the model by modulating the size of the embedding. Our observations indicate an inverted U-shaped relationship. Specifically, an escalation in the correlation score is noted when the embedding size is augmented from 256 to 1024. Conversely, a decrease in correlation is observed when the size is further increased from 1024 to 2048. The model attains the highest correlation score at an embedding size of 1024, while the lowest score is associated with an embedding size of 2048. It is important to note in this context that a larger embedding size does not necessarily equate to superior model performance.

Number of Heads. Within the framework of the Transformer architecture, the quantity of attention heads emerges as a pivotal parameter that could potentially influence the performance of

the model [97]. As such, we undertook an experiment wherein we varied the number of attention heads, while maintaining other hyperparameters at a constant level. Our observations revealed a negative correlation between the model's performance and the number of attention heads. More specifically, a marginal decrease in the model's performance was noted with an increase in the number of attention heads. The model achieved the highest scores with 4 attention heads, while the lowest score was observed when the model was equipped with 128 attention heads.

Patch Size. Within the visual encoder, the concept of patch size is employed to divide a frame into multiple patches for the purpose of linear projection. The dimensions of these patches play a crucial role in determining the extent of information that can be extracted, which, in turn, could influence the performance of the model. Our observations indicate a U-shaped correlation between the patch size and the model's performance. As the patch size transitions from 4×4 to 8×8, a marginal decline in the model's performance is noted. Conversely, an enhancement in the model's performance is observed when the patch size increases from 8×8 to 32×32. The highest correlation score is achieved with a patch size of 32×32, while the lowest score is associated with a patch size of 8×8.

Positional encoding. In the construction of the visual encoder, we have incorporated a hierarchical positional encoding scheme to infuse spatiotemporal information into the model. The efficacy of this approach is now under examination. A comparative analysis between the model performance devoid of positional encoding (as illustrated in row 16) and the model that employs positional encoding (as illustrated in row 1) reveals that the latter surpasses the former by a margin of 1%. Given that the rank correlation is computed for a substantial number of videos in the testing set, precisely 17,656, this enhancement in performance is statistically significant.

Number of Sampled Frames. Empirical evidence suggests that the performance of a

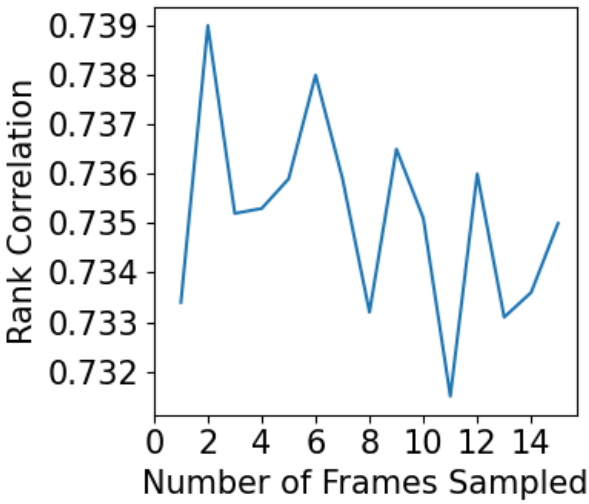


Figure 4.3: Spearman's Rank Correlations regarding the number of sampled frames.

model is contingent upon the number of frames sampled, as demonstrated in the ViViT study [9]. Consequently, we explore how the quantity of sampled frames impacts the model's performance, as illustrated in Figure 4.3. For the purpose of this investigation, we initially partitioned each video into 15 segments, sampling a single frame from each segment. The first n segments were sampled to represent the quantity of sampled frames. It was observed that the model's performance exhibited fluctuations in correlation with the increase in the number of frames. The model attained the highest correlation score when the number of frames was 2, while the lowest score was observed when the number of frames was 11.

4.7 Interpretability Analysis

The proposed model, referred to as the VAN, possesses a distinctive characteristic that facilitates the exploration of interrelationships among various modalities. This unique property is particularly significant as it provides a deeper understanding of how these modalities interact within the model.

To elaborate, the VAN model computes the product of the transformed query matrix $W^Q Q$ and the transpose of the transformed key matrix $W^K K$, resulting in an $n \times n$ matrix, denoted as R^h for each head h, where n represents the total number of modalities. The matrix R^h is of particular interest as it encapsulates the learned relationships among the different modalities. Each element in this matrix signifies the interaction strength between a pair of modalities, thereby providing a comprehensive view of the inter-modal relationships. Consequently, the contribution of a specific modality i for a particular head h, denoted as c_i^h, is quantified as the cumulative sum of its contributions to all modalities, inclusive of itself. This measure provides a comprehensive understanding of the role and influence of each modality in the context of each head within the multimodal learning framework. It is calculated as follows:

$$c_i^h = \sum_{j=1}^{n} R_{i,j}^h \tag{4.15}$$

In order to enhance the comparability of the contributions, we employ a softmax function. This function is utilized to transform the contributions into a probability distribution, thereby facilitating a more balanced comparison across different modalities. The resulting probability distribution, denoted as $\tilde{c}_i^h$, is calculated as follows:

$$\tilde{c}_i^h = \frac{e^{c_i^h}}{\sum_{i=1}^{n} e^{c_i^h}} \tag{4.16}$$

Therefore, the average contributions $\tilde{c}_i$ for a specific modality i is computed by averaging the contributions $\tilde{c}_i^h$ across all heads, denoted as $\tilde{c}_i$. This measure provides an aggregate understanding of the influence of each modality across all heads within the multimodal learning framework. It

is calculated as follows:

$$\tilde{c}_i = \frac{1}{n} \sum_{h=1}^{H} \tilde{c}_i^h \tag{4.17}$$

where H denotes the total number of heads.

We employ this methodology to investigate the contributions of all three modalities, namely vision, audio, and music network. The consolidated results of this investigation are depicted in Figure 4.4. This comprehensive analysis provides a holistic view of the relative importance and influence of each modality within the multimodal learning framework. It serves as a valuable resource for understanding the dynamics of multimodal interactions and their impact on the model's performance. We also include the contributions of each modality for each head in Appendix B.1, B.2, B.3 and B.4.

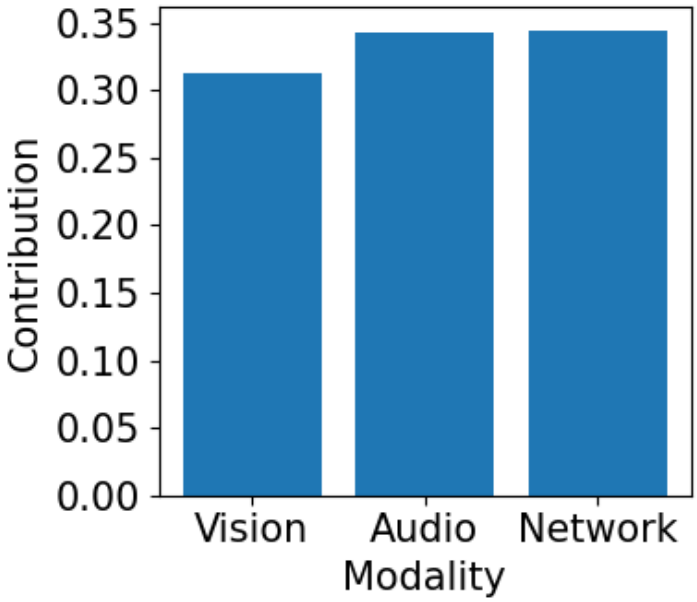


Figure 4.4: The average contributions of each modality.

As illustrated in Figure 4.4, the modalities of audio and network exhibit comparable average contributions to the model, both of which surpass that of vision. This observation suggests that

the vision modality holds less significance compared to the audio and network modalities within the context of this model.

This inference is further corroborated by the primary results obtained from the experiments. When the ViViT model is trained exclusively on vision data, it yields a Spearman's rank correlation of 0.4264. This performance is notably lower than when the model is trained on other modalities. In contrast, when the Audio-MLP and GAT models are trained separately using audio data and network data, they achieve a significantly higher rank correlation of 0.6289 and 0.6664 respectively. These results underscore the importance of the audio and network modalities in enhancing the model's performance and highlight the potential limitations of relying solely on the vision modality.

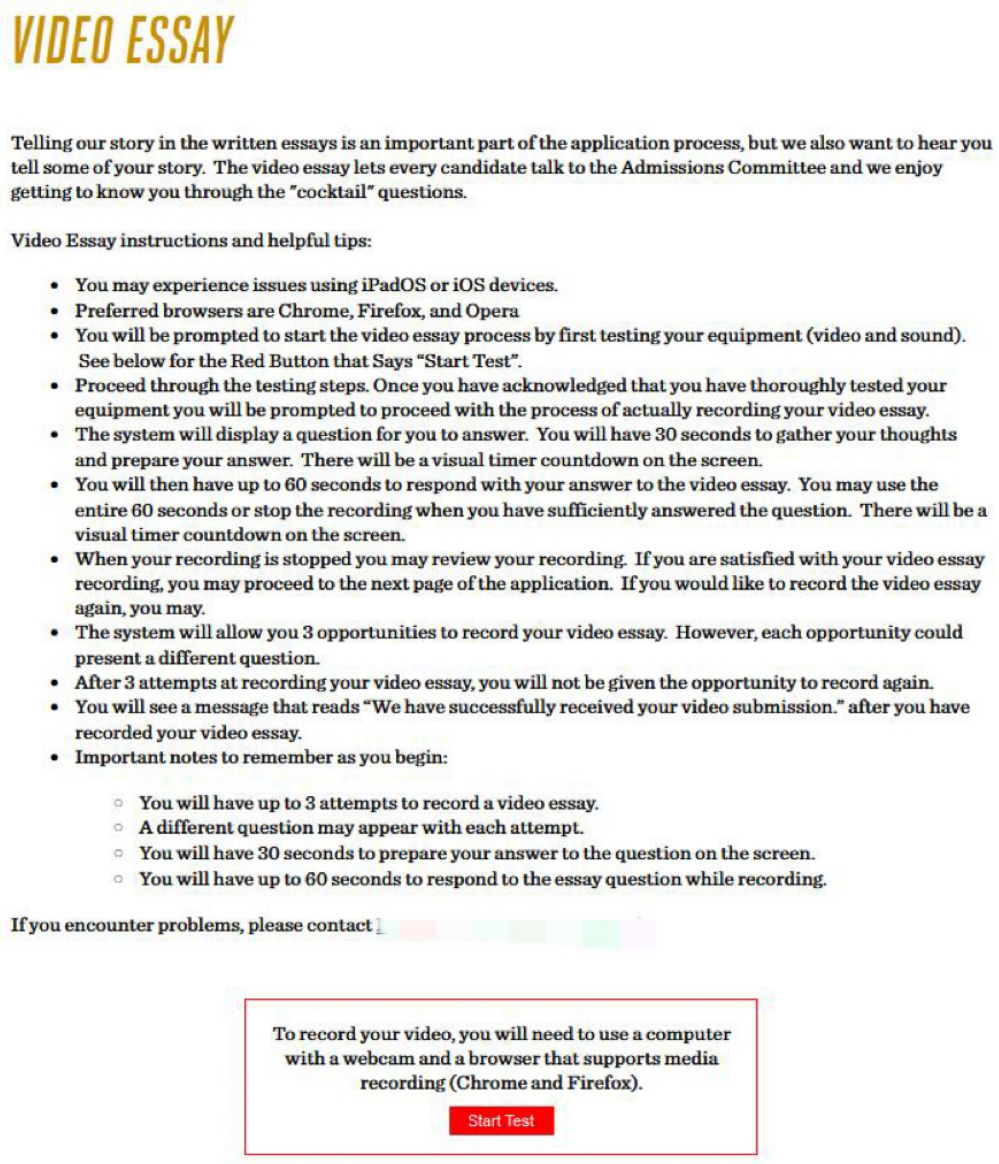
The analysis of the average contributions of each modality and the performance of the models trained on different modalities provide valuable insights into the relative importance of each modality in the multimodal learning process. These findings can guide future research in the field of multimodal learning and video popularity prediction.

4.8 Conclusion

In this study, we introduce an innovative and generalized multimodal framework specifically designed for the prediction of video popularity, namely VAN. This framework is unique in its ability to leverage a diverse set of multimodal features, which include but are not limited to visual cues, auditory signals, and most importantly, network characteristics. The primary strength of our proposed framework lies in its ability to leverage different categories of features. This leverage facilitates the efficient establishment of relationships between the multimodal features and the

corresponding popularity score of the video. In particular, it allows for a more comprehensive understanding of the modalities that contribute to video popularity. To further enhance the effectiveness of our framework, we have incorporated a Co-attention Transformer mechanism. This mechanism is instrumental in learning the relationships that exist among the three modalities - vision, audio, and network features. By doing so, it ensures that the interplay among these modalities is adequately captured and utilized in the prediction process. In addition to the above, we have also implemented a graph network pre-training strategy. The primary objective of this strategy is to improve the computational efficiency of our framework. By pre-training the graph network, we are able to expedite the learning process without compromising on the performance of the predictions. To validate the effectiveness of our proposed method, we have conducted both quantitative comparisons and ablation studies. The results from these rigorous evaluations provide compelling evidence of the superior performance of our method in predicting video popularity. We believe that our research contributes significantly to the field of multimodal learning and opens up new avenues for further exploration.

Appendix A: Video Essay Recording Page



VIDEO ESSAY

Telling our story in the written essays is an important part of the application process, but we also want to hear you tell some of your story. The video essay lets every candidate talk to the Admissions Committee and we enjoy getting to know you through the "cocktail" questions.

Video Essay instructions and helpful tips:

- You may experience issues using iPadOS or iOS devices.
- Preferred browsers are Chrome, Firefox, and Opera
- You will be prompted to start the video essay process by first testing your equipment (video and sound). See below for the Red Button that Says "Start Test".
- Proceed through the testing steps. Once you have acknowledged that you have thoroughly tested your equipment you will be prompted to proceed with the process of actually recording your video essay.
- The system will display a question for you to answer. You will have 30 seconds to gather your thoughts and prepare your answer. There will be a visual timer countdown on the screen.
- You will then have up to 60 seconds to respond with your answer to the video essay. You may use the entire 60 seconds or stop the recording when you have sufficiently answered the question. There will be a visual timer countdown on the screen.
- When your recording is stopped you may review your recording. If you are satisfied with your video essay recording, you may proceed to the next page of the application. If you would like to record the video essay again, you may.
- The system will allow you 3 opportunities to record your video essay. However, each opportunity could present a different question.
- After 3 attempts at recording your video essay, you will not be given the opportunity to record again.
- You will see a message that reads "We have successfully received your video submission." after you have recorded your video essay.
- Important notes to remember as you begin:
    - You will have up to 3 attempts to record a video essay.
    - A different question may appear with each attempt.
    - You will have 30 seconds to prepare your answer to the question on the screen.
    - You will have up to 60 seconds to respond to the essay question while recording.

If you encounter problems, please contact

To record your video, you will need to use a computer with a webcam and a browser that supports media recording (Chrome and Firefox).

Start Test

Figure A.1: A screenshot of the recording page for the video essay.

Appendix B: Additional Interpretability Analysis

The relative significance of each modality within each head is illustrated in Figures B.1, B.2, B.3, and B.4, respectively. In the context of Head 1 and Head 4, it is observed that audio information holds greater importance compared to both vision and network information, with network information marginally surpassing vision in terms of importance. Conversely, in Head 2 and Head 3, the importance of vision, audio, and network information is arranged in an arithmetic sequence. In this sequence, network information is deemed most important, while vision information is considered least important.

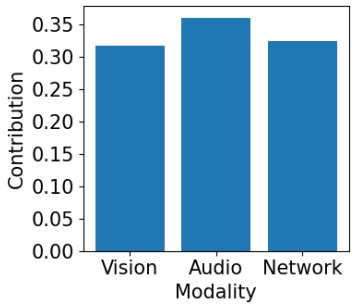


Figure B.1: The contributions of each modality for Head 1.

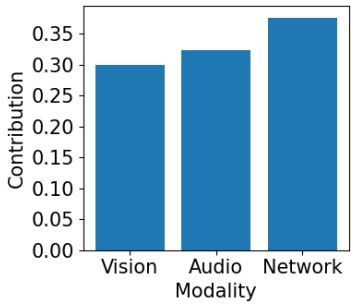


Figure B.2: The contributions of each modality for Head 2.

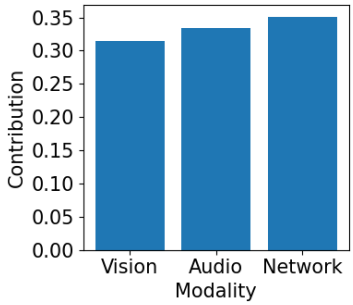


Figure B.3: The contributions of each modality for Head 3.

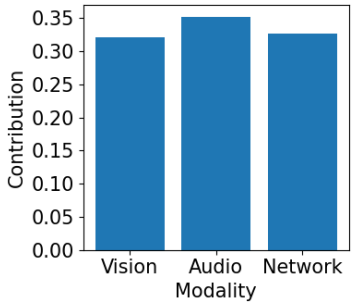


Figure B.4: The contributions of each modality for Head 4.

Bibliography

[1] Acar, E., Hopfgartner, F., and Albayrak, S. (2014). Understanding affective content of music videos through learned representations. In *MultiMedia Modeling: 20th Anniversary International Conference, MMM 2014, Dublin, Ireland, January 6-10, 2014, Proceedings, Part I 20*, pages 303–314. Springer.

[2] Adamopoulos, P., Ghose, A., and Todri, V. (2018). The impact of user personality traits on word of mouth: Text-mining social media platforms. *Information Systems Research*, 29(3):612–640.

[3] Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., and Gong, B. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text.

[4] Al-Samarraie, H., Eldenfria, A., and Dawoud, H. (2017). The impact of personality traits on users' information-seeking behavior. *Information Processing & Management*, 53(1):237–247.

[5] Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., and Zisserman, A. (2020). Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37.

[6] Allport, G. W. (1937). Personality: A psychological interpretation.

[7] Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., and Tran, D. (2020). Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770.

[8] Arandjelovic, R. and Zisserman, A. (2017). Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617.

[9] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846.

[10] Baveye, Y., Dellandréa, E., Chamaret, C., and Chen, L. (2015). Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55.

[11] Bechara, A. (2004). The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. *Brain and cognition*, 55(1):30–40.

[12] Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

[13] Berger, J. and Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2):192–205.

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

[15] Bielski, A. and Trzcinski, T. (2018). Pay attention to virality: understanding popularity of social media videos with the attention mechanism. *CoRR*, abs/1804.09949.

[16] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *ECCV 2020, Part I 16*, pages 213–229. Springer.

[17] Cartwright, M., Cramer, J., Salamon, J., and Bello, J. P. (2019). Tricycle: Audio representation learning from sensor network data using self-supervision. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 278–282. IEEE.

[18] Ceci, L. (2022). Number of live video viewers in the united states from 2019 to 2024. `https://www.statista.com/statistics/1284059/usa-live-video-viewership/`. Accessed: 7/5/2024.

[19] Ceci, L. (2024). Tiktok annual consumer spending worldwide from 2016 to 2023. `https://www.statista.com/statistics/1428578/global-tiktok-consumer-yearly-spending/`. Accessed: 7/5/2024.

[20] Chang, B., Zhu, H., Ge, Y., Chen, E., Xiong, H., and Tan, C. (2014). Predicting the popularity of online serials with autoregressive models. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 1339–1348, New York, NY, USA.

[21] Chen, J., Song, X., Nie, L., Wang, X., Zhang, H., and Chua, T.-S. (2016). Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, page 898–907, New York, NY, USA.

[22] Cheng, Y., Wang, R., Pan, Z., Feng, R., and Zhang, Y. (2020). *Look, Listen, and Attend: Co-Attention Network for Self-Supervised Audio-Visual Representation Learning*, page 3884–3892. Association for Computing Machinery, New York, NY, USA.

[23] Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y., and Lee, L.-S. (2016). Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. *arXiv preprint arXiv:1603.00982*.

[24] Cobb-Clark, D. A. and Schurer, S. (2012). The stability of big-five personality traits. *Economics Letters*, 115(1):11–15.

[25] Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., and McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*, 18(1):75.

[26] Crayne, M. P. and Medeiros, K. E. (2021). Making sense of crisis: Charismatic, ideological, and pragmatic leadership in response to covid-19. *American Psychologist*, 76(3):462.

[27] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

[28] Devaraj, S., Easley, R. F., and Crant, J. M. (2008). Research note—how does personality matter? relating the five-factor model to technology acceptance and use. *Information systems research*, 19(1):93–105.

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

[30] Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

[31] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[32] Ekman, P. et al. (1999). Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

[33] Gao, H., Kong, D., Lu, M., Bai, X., and Yang, J. (2018). Attention convolutional neural network for advertiser-level click-through rate forecasting. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1855–1864, Republic and Canton of Geneva, CHE.

[34] Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., and Misra, I. (2022). Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112.

[35] Goldberg, L. R. (1990). An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

[36] Güçlütürk, Y., Güçlü, U., van Gerven, M. A., and van Lier, R. (2016). Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In *ECCV 2016 Workshops, Part III 14*, pages 349–358.

[37] Hans, C., Suhartono, D., Andry, C., and Zamli, K. (2021). Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, 8(68).

[38] Hatfield, E., Cacioppo, J. T., and Rapson, R. L. (1993). Emotional contagion. *Current directions in psychological science*, 2(3):96–100.

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

[40] Hwang, S., Liu, X., and Srinivasan, K. (2021). Voice analytics of online influencers—soft selling in branded videos. *Available at SSRN 3773825*.

[41] Iashin, V. and Rahtu, E. (2020). A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *CoRR*, abs/2005.08271.

[42] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR.

[43] Kim, H., Lu, X., Costa, M., Kandemir, B., Adams, R. B., Li, J., Wang, J. Z., and Newman, M. G. (2018). Development and validation of image stimuli for emotion elicitation (isee): A novel affective pictorial system with test-retest repeatability. *Psychiatry Research*, 261:414–420.

[44] Kim, W., Son, B., and Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision.

[45] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[46] Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., and Prendinger, H. (2018). Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, 115:24–35.

[47] Kuhnke, F., Rumberg, L., and Ostermann, J. (2020). Two-stream aural-visual affect analysis in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 600–605. IEEE.

[48] Lampropoulos, G., Anastasiadis, T., Siakas, K., and Siakas, E. (2022). The impact of personality traits on social media use and engagement: An overview. *International Journal on Social and Education Sciences*, 4(1):34–51.

[49] LePine, J. A. and Van Dyne, L. (2001). Voice and cooperative behavior as contrasting forms of contextual performance: evidence of differential relationships with big five personality characteristics and cognitive ability. *Journal of applied psychology*, 86(2):326.

[50] Li, J., Shomer, H., Mao, H., Zeng, S., Ma, Y., Shah, N., Tang, J., and Yin, D. (2023). Evaluating graph neural networks for link prediction: Current pitfalls and new benchmarking.

[51] Liu, L., Preotiuc-Pietro, D., Samani, Z. R., Moghaddam, M. E., and Ungar, L. (2016). Analyzing personality through social media profile picture choice. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 211–220.

[52] Liu, W., Qiu, J.-L., Zheng, W.-L., and Lu, B.-L. (2021a). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729.

[53] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

[54] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. (2022). Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211.

[55] Lu, S., Xiao, L., and Ding, M. (2016). A video-based automated recommender (var) system for garments. *Marketing Science*, 35(3):484–510.

[56] Luo, C., Jiang, Z. J., Li, X., Yi, C., and Tucker, C. (0). Choosing to discover the unknown: The effects of choice on user attention to online video advertising. *Management Science*, 0(0):null.

[57] Lynn, V., Balasubramanian, N., and Schwartz, H. A. (2020). Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5306–5316.

[58] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

[59] McRae, K., Misra, S., Prasad, A. K., Pereira, S. C., and Gross, J. J. (2012). Bottom-up and top-down emotion generation: implications for emotion regulation. *Social cognitive and affective neuroscience*, 7(3):253–262.

[60] Mitra, V. and Wang, C.-J. (2008). Content based audio classification: a neural network approach. *Soft Computing*, 12:639–646.

[61] Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of experimental psychology*, 64(5):482.

[62] Nadkarni, S. and Herrmann, P. (2010). Ceo personality, strategic flexibility, and firm performance: The case of the indian business process outsourcing industry. *Academy of Management Journal*, 53(5):1050–1073.

[63] Naumann, L. P., Vazire, S., Rentfrow, P. J., and Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and social psychology bulletin*, 35(12):1661–1671.

[64] Ou, N., Yu, L., Li, H., Du, Q., Xiang, J., and Gong, W. (2022). Mtaf: Shopping guide micro-videos popularity prediction using multimodal and temporal attention fusion approach. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4543–4547.

[65] Panksepp, J. (2004). *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press.

[66] Park, B., Tsai, J. L., Chim, L., Blevins, E., and Knutson, B. (2016). Neural evidence for cultural differences in the valuation of positive facial expressions. *Social Cognitive and Affective Neuroscience*, 11(2):243–252.

[67] Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., Baró, X., Escalante, H. J., and Escalera, S. (2016). Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *ECCV 2016 Workshops, Part III 14*, pages 400–418. Springer.

[68] Poria, S., Chaturvedi, I., Cambria, E., and Hussain, A. (2016). Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 439–448.

[69] Raab, M., Schlauderer, S., Overhage, S., and Friedrich, T. (2020). More than a feeling: Investigating the contagious effect of facial emotional expressions on investment decisions in reward-based crowdfunding. *Decision Support Systems*, 135:113326.

[70] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

[71] Rao, K. S. and Manjunath, K. (2017). *Speech recognition using articulatory and excitation source features*. Springer.

[72] Riaz, M. N., Riaz, M. A., and Batool, N. (2012). Personality types as predictors of decision making styles. *Journal of Behavioural Sciences*, 22(2).

[73] Russell, J. A. (1980a). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

[74] Russell, J. A. (1980b). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

[75] Sarangi, S., Sahidullah, M., and Saha, G. (2020). Optimization of data-driven filterbank for automatic speaker verification. *Digital Signal Processing*, 104:102795.

[76] Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology*, 8(4):467–487.

[77] Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition & Emotion*, 14(4):433–440.

[78] Shepherd, J. (2024). 30 vital video marketing statistics you need to know in 2024. `https://thesocialshepherd.com/blog/video-marketing-statistics`. Accessed: 7/5/2024.

[79] Singh, C. (2023). 100+ surprising video marketing statistics you should know in 2023.

[80] Stieglitz, S. and Dang-Xuan, L. (2013). Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4):217–248.

[81] Su, M.-H., Wu, C.-H., Huang, K.-Y., and Hong, Q.-B. (2018). Lstm-based text emotion recognition using semantic and emotional word vectors. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6. IEEE.

[82] Subramaniam, A., Patel, V., Mishra, A., Balasubramanian, P., and Mittal, A. (2016). Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *ECCV 2016 Workshops, Part III 14*, pages 337–348. Springer.

[83] Suman, C., Saha, S., Gupta, A., Pandey, S. K., and Bhattacharyya, P. (2022). A multi-modal personality prediction system. *Knowledge-Based Systems*, 236:107715.

[84] Sun, M. and Zhang, K. (2023). Multimodal co-attention transformer for video-based personality understanding. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1450–1459.

[85] Sun, M. and Zhang, K. (2024). Sec2sec co-attention transformer for video-based apparent affective prediction. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8255–8259.

[86] Sun, X., Liu, B., Cao, J., Luo, J., and Shen, X. (2018). Who am i? personality detection based on deep learning for texts. In *2018 IEEE international conference on communications (ICC)*, pages 1–6. IEEE.

[87] Sun, Z., Sarma, P., Sethares, W., and Liang, Y. (2019). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis.

[88] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks.

[89] Tagliasacchi, M., Gfeller, B., Quitry, F. d. C., and Roblek, D. (2019). Self-supervised audio representation learning for mobile devices. *arXiv preprint arXiv:1905.11796*.

[90] Tang, S., Li, Q., Ma, X., Gao, C., Wang, D., Jiang, Y., Ma, Q., Zhang, A., and Chen, H. (2022a). Knowledge-based temporal fusion network for interpretable online video popularity prediction. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2879–2887, New York, NY, USA.

[91] Tang, Z., Cho, J., Nie, Y., and Bansal, M. (2022b). Tvlt: Textless vision-language transformer.

[92] Team, B. (2024). Tiktok statistics you need to know. `https://backlinko.com/tiktok-users`. Accessed: 7/5/2024.

[93] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[94] Trosten, D. J., Løkse, S., Jenssen, R., and Kampffmeyer, M. (2021). Reconsidering representation alignment for multi-view clustering.

[95] Trzcinski, T., Andruszkiewicz, P., Bochenski, T., and Rokita, P. (2017). Recurrent neural networks for online video popularity prediction. *CoRR*, abs/1707.06807.

[96] Tsao, W.-C. and Chang, H.-R. (2010). Exploring the impact of personality traits on online shopping behavior. *African journal of business management*, 4(9):1800.

[97] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

[98] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks.

[99] Wang, J., Wang, Y., Weng, N., Chai, T., Li, A., Zhang, F., and Yu, S. (2021). Will you ever become popular? learning to predict virality of dance clips.

[100] Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016). Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

[101] Wei, X., Zhang, C., Zhang, H., and Wu, J. (2018). Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, 9(03):303–315.

[102] Wei, X.-S., Zhang, C.-L., Zhang, H., and Wu, J. (2017). Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, 9(3):303–315.

[103] Weinel, J. (2018). *Inner sound: altered states of consciousness in electronic music and audio-visual media*. Oxford University Press.

[104] Wikipedia (2024). Emotion. `https://en.wikipedia.org/wiki/Emotion`. Accessed: 7/5/2024.

[105] Witt, T. (2023). Social media video statistics marketers need to know for 2023.

[106] Wyzowl (2024). Video marketing statistics 2024. `https://www.wyzowl.com/video-marketing-statistics/`. Accessed: 7/5/2024.

[107] Xie, J., Zhu, Y., Zhang, Z., Peng, J., Yi, J., Hu, Y., Liu, H., and Chen, Z. (2020). A multimodal variational encoder-decoder framework for micro-video popularity prediction. In *Proceedings of The Web Conference 2020*, WWW '20, page 2542–2548, New York, NY, USA.

[108] Xue, D., Wu, L., Hong, Z., Guo, S., Gao, L., Wu, Z., Zhong, X., and Sun, J. (2018). Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, 48:4232–4246.

[109] Yang, K., Lau, R. Y., and Abbasi, A. (2023). Getting personal: a deep learning artifact for text-based measurement of personality. *Information Systems Research*, 34(1):194–222.

[110] Yin, D., Bond, S. D., and Zhang, H. (2014). Anxious or angry? effects of discrete emotions on the perceived helpfulness of online reviews. *MIS quarterly*, 38(2):539–560.

[111] Yu, J. and Markov, K. (2017). Deep learning based personality recognition from facebook status updates. In *2017 IEEE 8th international conference on awareness science and technology (iCAST)*, pages 383–387. IEEE.

[112] Zhang, Y., Li, P., Zhang, Z., Zhang, C., Wang, W., Ning, Y., and Lian, B. (2020). Graphinf: A gcn-based popularity prediction system for short video networks. page 61–76, Berlin, Heidelberg. Springer-Verlag.

[113] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H., et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of CVPR*, pages 6881–6890.

[114] Zhou, R. and Shen, Y.-D. (2020). End-to-end adversarial-attention network for multi-modal clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[115] Zhu, Y., Xie, J., and Chen, Z. (2020). Predicting the popularity of micro-videos with multimodal variational encoder-decoder framework. *CoRR*, abs/2003.12724.