EMOLIS App and Dataset to Find Emotionally Close Cartoons

Soëlie Lerch, Patrice Bellot, Emmanuel Bruno, Elisabeth Murisasco
Aix Marseille Univ, Université de Toulon, CNRS, LIS
Toulon - Marseille, France
{firstname.lastname}@lis-lab.fr

Abstract

We propose EMOLIS Dataset that contains annotated emotional transcripts of scenes from Walt Disney cartoons at the same time as physiological signals from spectators (breathing, ECG, eye movements). The dataset is used in EMOLIS App, our second proposal. EMOLIS App allows to display the identified emotions while a cartoon is playing and suggests emotionally comparable videos. We propose to estimate an emotional distance between videos using multimodal neural representations (text, audio, video) that also combine physiological signals. This enables personalized results that can be used for cognitive therapies focusing on awareness of felt emotions. The dataset is designed to be suitable for all audiences and autistic people who have difficulties to recognize and express emotions.

Keywords: emotion analysis, video retrieval, interactive retrieval, autism spectrum disorder (ASD)

1. Introduction

When we are interested in the automated emotions qualification conveyed by content, the question is to what extent we can be satisfied with the content only, independently from its reception and conscious feelings. People with autism spectrum disorders (ASD) often have difficulty identifying the emotions they feel, so it's worth taking an interest in recognizing them. Using emotionally comparable cartoon scenes can help with this training, but you need to be able to find them. That is how this article deals with automated emotions analysis from the point of view of emitted emotions (here by cartoons) and felt emotions (here captured by physiological signals). Cartoon scenes include visual, audio and spoken signals which communicate emotions and generate conscious or unconscious physiological reactions in viewers. We present EMOLIS DataSet, multimodal and physiological data (Section 3), and EMOLIS App (Section 4), which estimates the emotions conveyed by a video and suggests related emotional content. The dataset and source code are available online (GitHub, Zenodo).

2. Background and motivations

Automated emotion analysis is deeply described in (Guo et al., 2019), (Sharma and Dhall, 2021) and has been the subject of several tasks in Semeval, Mediaeval, Evalita, EmotiW[1]. Most methods exploit supervised approaches applied to annotated datasets with emotions. Multimodal neural architectures (CNN (LeCun et al., 1998), GAN (Goodfellow et al., 2014), LSTM (Hochreiter and Schmidhuber, 1997), GRU (Cho et al., 2014)) achieve accuracy and F-measure between 66 and 78% (Nie et al., 2021), (Huan et al., 2021) for identifying emotions. But considering emotions from only video content does not make it easy to consider the variability of individual feelings and therefore to learn models adapted to viewers. On the other hand, the felt emotions can be well detected by physiological signals (Muszynski, 2018). These different results motivate the search for approaches integrating these signals as a complement to content. Zhou et al. (2023) summarizes all the multimodal recommendation and fusion methods for this task but to our knowledge, physiological signals have not been used for video recommendation yet, as they have been used for music (Liu and Hu, 2020).

In this work, we aim to target a young people with autism spectrum disorders who often have difficulties to recognize and verbalize the emotions they feel. Cognitive therapies focusing on emotion recognizing and learning employ pictures, videos or songs (see for example https://modelmekids.com/autism-emotions/). Computerized training programs have been proposed (Farashi et al., 2022) but they are based on a limited corpus of videos, which are expensive and complex to build. One of our ideas is to make it possible to exploit larger and unannotated corpora by using emotional signatures that can be calculated on any video.

We wish to use cartoons because their scenes are generally less rich in disruptive objects, with simplified facial features (Atherton and Cross, 2018) and a low speed of graphic variation: rapid variations can interfere with emotional reception (Tardif et al., 2017). While many studies of emotional reception in ASD children have been based on still images or photographs, there are few on cartoon images (Jain et al., 2021) and even less on animated cartoons and movies. Unfortunately, there are not

[1] See semeval.github.io, multimediaeval.github.io, www.evalita.it, sites.google.com/view/emotiw2020

many annotated emotional datasets either. Liris-Accede (Baveye et al., 2015) contains English film scenes and a few in French. Cartoon scenes are present, but with little dialogue and the scenes are often long, some are not suitable for children. Amigos (Santamaria-Granados et al., 2018) contains English short videos and physiological signals, but is not suitable for a young audience either. The MELD dataset (**?**) is suitable for fairly young viewers (Friends TV show) and annotated with six Ekman's emotions (joy, anger, fear, disgust, sadness, surprise) and polarity. But, it does not present a wide diversity of emotions: joy is largely dominant.

3. EMOLIS Dataset

3.1. Modalities extracted from the video

EMOLIS Dataset is composed of sixty-two cartoon scenes from Walt-Disney films (between seven and nine scenes per film) adapted for a young audience and available in some fifty languages, including text, image and audio modalities. For each film, the scenes are delimited by a change of decor or subject of conversation between characters. These scenes have been chosen based on their portrayal of emotions, specially focusing on scenes that exhibit a higher degree of dominance within the spectrum of Ekman emotion.

We manually cut the scenes into sequences: a sequence corresponds to a turn of speech, or to a silence of at least one second. This allows to predict emotions in real time, by character and by moment of silence (facial expressions). The total duration of the EMOLIS corpus is two hours, eight minutes and thirty-eight seconds.

Speaker	t_1	t_2	t_3	t_4	t_5	t_6	t_7
Elsa	A	B	C	D	_	_	_
Anna	_	_	V	W	X	Y	_

Speaker	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}
Elsa	_	E	F	_	G	_	H
Anna	_	_	Z	_	_	_	_

Table 1: Temporal representation of speech turns. _ represents silence, and letters sentences.

The transcripts of the scenes were obtained from subtitles from online sites. They are composed of a succession of dialogue with the name of the character speaking. Transcripts are wrapped from subtitles and subdivided into speech turns. For each film, we built a CSV file containing one line per sequence (speech turn or silence), with scene number, turn number, the character's name, and the start and end time markers in the video (based by subtitles).

Table 3 shows an example of a data file that allows to align videos (Table 1) and transcripts (Table 2).

line #	speaker	s_1	s_2	s_3	s_4	start time	end time
1	Elsa	A	B	C	D	t1	t4
2	Anna	V	W	X	Y	t3	t6
3	Elsa	E	F			t9	t10
4	Anna	Z				t10	t10
5	Elsa	G				t12	t12
6	Elsa	H				t14	t14

Table 2: Transcript file to be aligned with videos.

At each moment t (Table 1), a sentence is spoken and some other characters may speak at the same time (for example at the moment t_3).

	scene	utterance	speaker	start time	end time
1	6	1	Elsa	$t1$	$t4$
2	6	2	Anna	$t3$	$t6$
3	6	3	–	$t7$	$t8$
4	6	4	Elsa	$t9$	$t10$
5	6	5	Anna	$t10$	$t10$
6	6	6	–	$t11$	$t11$
7	6	7	Elsa	$t12$	$t12$
8	6	8	–	$t13$	$t13$
9	6	9	Elsa	$t14$	$t14$

Table 3: Data file for aligning transcripts and videos. Line 1 indicates that Elsa says four sentences, from instant t_1 to instant t_4. Transcript is on line 1 of Table 2, with s_i the i^{th} sentence. Line 3 represents a silence (no character speaking) from instant t_7 to t_8. Lines 1-2 and 4-5 show Elsa and Anna speaking at the same time at instants t_3, t_4 and t_{10}. Lines 7 to 9 correspond to an Elsa's monologue interrupted by silence.

We share EMOLIS Dataset on Zenodo.org. In particular, it contains the URLs of the original subtitles, the scripts for extracting and formatting data, and CSV files containing data for aligning the videos and the transcripts.

3.2. Collection of physiological signals

EMOLIS Dataset contains also physiological signal measures (breathing, electrocardiogram, eye movements) which can be used to refine and personalize the emotional signatures of the videos. The distributed data corresponds to only a few people but the idea is not to claim generic data, but to show that each viewer can build his or her own personalized model. These data are obtained as presented in Figure 1. The viewer sits in front of

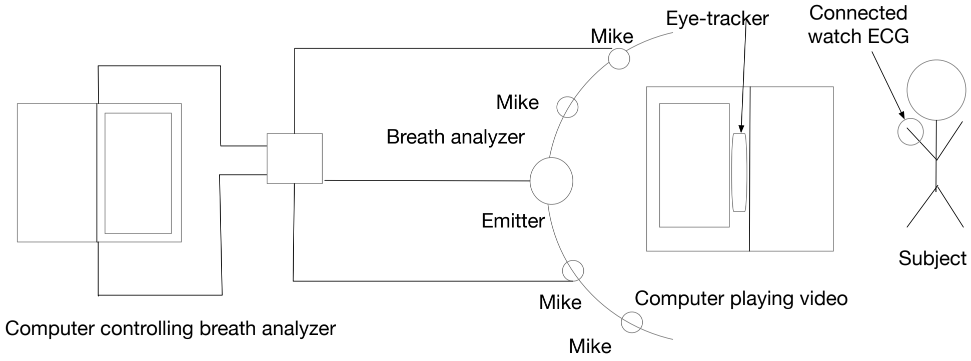

Figure 1: Hardware installation for physiological data collection and video viewing.

a computer which plays the film and records both eye movements by means of a Tobii Pro Nano eye tracker at the bottom of the screen and facial expressions from the integrated webcam. The viewer wears an Apple Watch which records electrocardiogram. Finally, an home-made breath analyzer positioned around the computer detects breaths by ultra-sound: frequency domains representative of exhalations are recorded. After the signals have been recorded, the viewer annotates the emotions he or she felt during the different scenes.

The data collected by the eye tracker are the date and time, the size of the two pupils and the corresponding position on the screen for each eye. From this data, we calculated the values of the descriptors which are generally used for emotion recognition: pupil size (Moharana and Das, 2021), eye position, duration and number of blinks (Babu and Lahiri, 2020), fixation points (Deng and Gu, 2020), fixation duration(Martínez-Velázquez et al., 2020), number and duration of saccades: this corresponds to the rapid movement of the eyes between two positions (Schmidt et al., 2012). In addition to these classic descriptors, we added the angle of rotation between the starting point and two successive saccades to be able to take account to the change of gaze direction.

The ECG device in the watch records variations in the electrical potential of the myocardial membrane, a signal in volts as a function of time. We extract a pre-trained representation from the filtered and normalized signal using a neural architecture and a pre-trained model from (Sarkar and Etemad, 2020).

3.3. Annotation of felt emotions

Seven subjects, including the authors, have viewed all the scenes to annotate emotions. For each scene, each annotator has to identify the chronology of felt emotions and count them. Then, the annotator selects the most representative emotion(s) or indicates "other"(o) or "no emotion"(none). Annotation time is around 5 hrs 20 mins for one annotator, corresponding to 2 hrs 8 mins of video.

The viewer's annotations are in CSV files. Lines contain scene number, frequency of emotions, the dominant emotion and the sequence of emotional appearance (Table 4).

3.4. Inter-annotator agreements.

The annotators are aged from 12 to 64 and from a variety of backgrounds and levels of education. Two of them have Asperger's autism (aged 23 and 27). From their annotations, we count the number of times where emotions are present in a scene. We take a majority vote among all the annotators. Twice, annotators were unable to choose a dominant emotion, then two emotions were retained for two scenes.

Table 5 shows the distribution of the major emotions in the scenes. Emotions are well distributed among joy (j), fear (f) and sadness (sa), but not for anger (a), surprise (su) and disgust (d). Fleiss Kappa scores (Fleiss, 1971) between annotators are presented for each emotion (Table 6). It measures agreement between the annotators, unlike Cohen Kappa (Cohen, 1960), which measures agreements in pairs.

Kappa (Table 6) is medium for joy (0.4), anger and fear (0.5), weak for disgust (0.3) and surprise (0.1) and significant for sadness (0.6). In comparison, the Kappa score for the MELD dataset (**?**) of TV series is 0.43.

The impact of socio-cultural and personal factors is known to be significant (Krohne, 2003). In our case, four annotators are members of the same family, and for them Cohen kappa score calculated two by two is significant (0.61 to 0.78) for fear and medium to significant (0.53 to 0.74) for sadness. Whereas for those who do not belong to the same family, the scores vary widely from weak to significant (0.2 to 0.58) for fear and (0.35 to 0.65) for sadness. Moreover, the variability among annotators may be due to the difficulty of naming emotions for Asperger autistic persons and/or the difficulty in choosing the

film	scene	occurrences of emotions							major emotion								sequence
		a	d	j	f	su	sa	o	a	d	j	f	su	sa	o	n	
Mulan	1	0	1	6	0	0	1	0	0	0	1	0	0	0	0	0	$j\ j\ j\ d\ j\ j\ j\ sa$

Table 4: Number of occurrences of each emotion and the most representative emotion, joy (j), experienced by one annotator of Scene 1 from Mulan. Sequence indicates that the annotator first felt joy (j) 3 times, then disgust (d), then joy again 3 times and lastly sadness (sa). The most representative emotion is joy. Anger (a), fear (f), surprise (su) and other (o) were not felt. (n) corresponds to "no emotion felt".

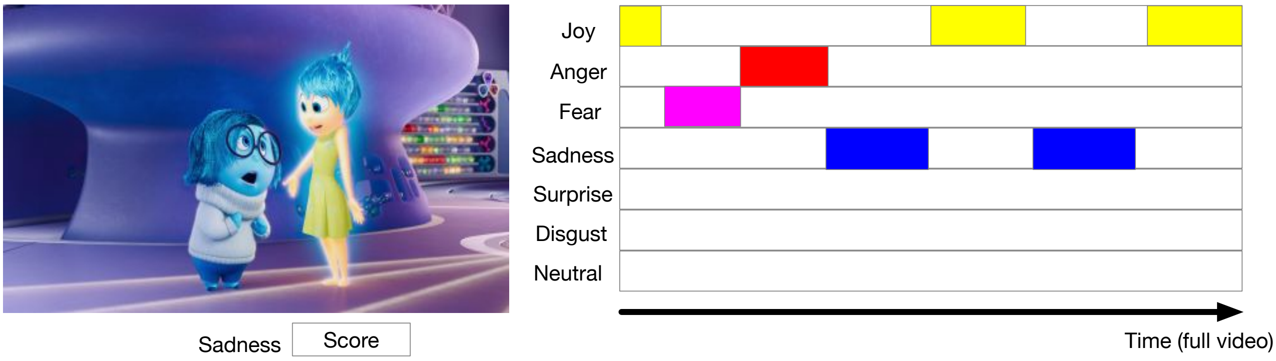

Figure 2: Real-time display of the emotions conveyed by the current scene (left) and full video (right).

major emotion	j	a	f	sa	d	su	o
# of scenes	19	7	17	16	5	4	0

Table 5: Major emotions distribution in the scenes.

Fusion	j	a	f	sa	d	su	o
Kappa	0.4	0.4	0.5	0.6	0.3	0.1	0

Table 6: Fleiss Kappa scores between annotators (of the most representative emotions).

most representative emotions.

All this highlights the difficulty of annotating with felt emotions. The highly subjective nature of the task requires having a very diverse and large panel of annotators or categorizing them into groups based on criteria to be determined. This lends further support to the aim of comparing emotional signatures rather than trying to name the emotions themselves.

Our system has not yet been tested with children with ASD but only with adults. However, these adults underwent childhood training in the form of a game to practice identifying their emotions and recognizing emotions expressed by characters. The prototype currently in development aims to be a dynamic version of this training game (capable of working with new videos). Furthermore, ASD adults are usually very sensitive and can feel emotions as strong as children. The proposed solution can be helpful for them too. Contacts are underway with psychiatrists and therapists who are interested in experimenting with this approach with certain patients as a means to support understanding and decoding of their emotions.

4. EMOLIS App

Now, we present EMOLIS App, which leverages the previously described data. The source code of EMOLIS App is distributed on Github. It allows real-time display of emotions conveyed by a user-selected video and suggests similar emotional scenes based on content and physiological signals.

Therefore, the system's goal is twofold: to recommend scenes to a viewer that evoke emotions similar to those experienced while watching a given video and to monitor in real-time the emotions expressed by the characters. Children with ASD (Autism Spectrum Disorder) struggle to name their own emotions and those expressed by their interlocutors. The developed tool would help them learn to identify these emotions.

(Figure 2).

The emotions expressed by the character are used for real-time emotion prediction (categorization). They are estimated using a neural architecture derived from (Siriwardhana et al., 2020). It uses pretrained representations for audio (Wav2vec), text (Roberta) and image (Fabnet Video) modalities. To obtain representations (CLS tokens) for audio and image, fine-tuned unimodal transformers are used to identify emotions from audio and image representations. The audio models we use are trained from spoken voice. Because audio from movies contains sound and music too, we isolated voice by means of Spleeter (Hennequin et al., 2020).

In order to cross modalities, bimodal transformers are applied resulting in six bi-modal output representations. After applying a Hadamar product to each pair, we concatenate the results to obtain a single representation for each sequence. After applying the Hadamard product to each pair, we concatenate the three representations (text, facial expression, voice) to create a unified representation for each sequence, encompassing all three modalities. Subsequently, a model is trained using two linear layers and dropout to produce a tensor with a size of seven (representing each Ekman emotion). A Softmax function is then applied to derive probabilities for each emotion. The emotion with the highest probability is considered the predicted emotion. We achieved a 61% F-score in emotion prediction on the MELD dataset, which is comparable to the results mentioned in (Siriwardhana et al., 2020).

To calculate an "emotional" distance between two videos, we use, as in (Rouabhia and Tebbikh, 2010), the distance derived from the Frobenius matrix norm on the matrix representations associated with each type of physiological signal, as well as on the multimodal representation of the video content (4 matrices). The emotions felt by the viewer (represented by a signature derived from physiological signals) are used. The emotional distance between two scenes is then seen as the sum of these distances. So, EmolisApp can retrieve videos with the closest emotional representations. The user can evaluate the result by scoring it (from 1 to 4). We must improve the evaluation of our system, exploring two potential avenues for assessment— annotations provided by annotators and user ratings. The key question for collecting the ratings is: does the recommended video meet the user's expectations? A user could be a person training to recognize emotions or, for example, a psychiatric or psychological doctor using the application with his or her practice. Moving forward, our aim is to enhance the system if it fails to meet these expectations. In a future version, the ratings could fed back into the system but this has not yet been implemented.

5. Conclusion

We presented a new dataset composed of annotated cartoons and an application that retrieves emotionally similar videos. There are several innovations in our proposal. First, we developed EMOLIS Dataset, a new multimodal annotated dataset based on popular cartoons (transcripts aligned with video) associated to physiological data. Secondly, considering physiological signals to suggest emotionally similar videos based on the viewer's feelings (physiological data) is also a novel aspect. Lastly, we employed a state-of-the-art neural architecture to represent the cartoons and obtain emotional signatures that allow to estimate emotional distances between them.

In future, we will use EMOLIS App in the context of theory of mind disorders and for emotional training of young people with autism spectrum disorder.

6. References

Gray Atherton and Liam Cross. 2018. Seeing more than human: autism and anthropomorphic theory of mind. *Frontiers in psychology*, 9:528.

Pradeep Raj Krishnappa Babu and Uttama Lahiri. 2020. Classification approach for understanding implications of emotions using eye-gaze. *Journal of Ambient Intelligence and Humanized Computing*, 11(7):2701–2713.

Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. 2015. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Mingming Deng and Xiuzhu Gu. 2020. Information acquisition, emotion experience and behaviour intention during online shopping: an eye-tracking study. *Behaviour & Information Technology*, pages 1–11.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Sajjad Farashi, Saeid Bashirian, Ensiyeh Jenabi, and Katayoon Razjouyan. 2022. Effectiveness of virtual reality and computerized training programs for enhancing emotion recognition in people with autism spectrum disorder: a systematic review and meta-analysis. *International Journal of Developmental Disabilities*, pages 1–17.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394.

Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154. Deezer Research.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ruo-Hong Huan, Jia Shu, Sheng-Lin Bao, Rong-Hua Liang, Peng Chen, and Kai-Kai Chi. 2021. Video multimodal emotion recognition based on bi-gru and attention fusion. *Multimedia Tools and Applications*, 80(6):8213–8240.

Nikita Jain, Vedika Gupta, Shubham Shubham, Agam Madan, Ankit Chaudhary, and KC Santosh. 2021. Understanding cartoon emotion using integrated deep neural network on large dataset. *Neural Computing and Applications*, pages 1–21.

Heinz W Krohne. 2003. Individual differences in emotional reactions and coping.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Ruilun Liu and Xiao Hu. 2020. A multimodal music recommendation system with listeners' personality and physiological signals. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 357–360.

Eduardo S Martínez-Velázquez, Alma L Ahuatzin González, Yaira Chamorro, and Henrique Sequeira. 2020. The influence of empathy trait and gender on empathic responses. a study with dynamic emotional stimulus and eye movement recordings. *Frontiers in psychology*, 11:23.

Laxmipriya Moharana and Niva Das. 2021. Analysis of pupil dilation on different emotional states by using computer vision algorithms. In *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON)*, pages 1–6. IEEE.

Michal Muszynski. 2018. *Recognizing film aesthetics, spectators' affect and aesthetic emotions from multimodal signals*. Ph.D. thesis, University of Geneva.

Weizhi Nie, Yan Yan, Dan Song, and Kun Wang. 2021. Multi-modal feature fusion based on multi-layers lstm for video emotion recognition. *Multimedia Tools and Applications*, 80(11):16205–16214.

Chahrazed Rouabhia and Hicham Tebbikh. 2010. Mesure de similarité pondérée dans l'espace 2d: Application à la reconnaissance de visages. In *CORIA*, pages 373–385.

Luz Santamaria-Granados, Mario Munoz-Organero, Gustavo Ramirez-Gonzalez, Enas Abdulhay, and NJIA Arunkumar. 2018. Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos). *IEEE Access*, 7:57–67.

Pritam Sarkar and Ali Etemad. 2020. Self-supervised ecg representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 13(3):1541–1554.

Lisette J Schmidt, Artem V Belopolsky, and Jan Theeuwes. 2012. The presence of threat affects saccade trajectories. *Visual Cognition*, 20(3):284–299.

Garima Sharma and Abhinav Dhall. 2021. A survey on automatic multimodal emotion recognition in the wild. In *Advances in Data Science: Methodologies and Applications*, pages 35–64. Springer.

Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. 2020. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access*, 8:176274–176285.

Carole Tardif, Laura Latzko, Thomas Arciszewski, and Bruno Gepner. 2017. Reducing information's speed improves verbal cognition and behavior in autism: A 2-cases report. *Pediatrics*, 139(6).

Hongyu Zhou, Xin Zhou, Zhiwei Zeng, Lingzi Zhang, and Zhiqi Shen. 2023. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473*.