VACD: A Video Affective Dataset with Content Descriptions (Appendix)

Jinqiao Lu*
University of Science and Technology of China
Hefei, China
ljq18317370153@mail.ustc.edu.cn

Jingtian Li
University of Science and Technology of China
Hefei, China
lijingtian@mail.ustc.edu.cn

Yadong Liu
University of Science and Technology of China
Hefei, China
yadongliu@mail.ustc.edu.cn

Caichao Zhang
University of Science and Technology of China
Hefei, China
zcc1101@mail.ustc.edu.cn

Zhenjie Liu
University of Science and Technology of China
Hefei, China
liuzhenjie@mail.ustc.edu.cn

Haofan Zhang
University of Science and Technology of China
Hefei, China
haofan0202@mail.ustc.edu.cn

Jiale Huang
University of Science and Technology of China
Hefei, China
huangjiale@mail.ustc.edu.cn

Mintao Zhang
University of Science and Technology of China
Hefei, China
ustccs_zmt@mail.ustc.edu.cn

A Statistical Analysis of the Dataset

A.1 Statistical Analysis of the Dataset

The histogram of video clip durations in the dataset is shown in Figure 4. To assess the emotional richness of VACD, we conducted statistical analyses using the majority-vote annotations from the three annotators for each movie. We first examined the number of emotion labels assigned to each clip. We found that 58% of the clips were annotated with at least two emotion labels, as shown in the Figure 3. We then analyzed the distributions of valence and arousal values in the majority-vote annotations. As shown in the Figure 5, results show that 60.9% of the clips have a valence value of either -1 or 1, indicating that most clips exhibit clear affective polarity. In addition, 77.2% of the clips are annotated with arousal values of 1 or 2, suggesting that a large proportion of clips convey relatively strong emotional intensity. It should be noted that the statistical results presented in Figures 3 and 5 were calculated based on the modal annotation results across the three annotators per movie. To further evaluate the dataset's representativeness, we analyzed the demographics of the 65 annotators who contributed to the labeling process. The annotators consisted of 36 males and 29 females (Figure 2). Their age distribution ranged from 20 to 50 (Figure 1).

*Corresponding author, ljq18317370153@mail.ustc.edu.cn

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM'26, Rio de Janeiro, Brazil

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

A.2 Inter-Annotator Consistency Analysis

To demonstrate the reliability and representativeness of the labels in the dataset, we conducted a consistency analysis of the annotations provided by all annotators. For this purpose, we utilized the mode value M of the annotations from three annotators assigned to each movie during dataset construction. Since the dataset includes binary classification for the ten basic emotions and ternary classification for valence and arousal, different consistency evaluation metrics were adopted accordingly.

For the ten basic emotions annotated using binary classification, we adopted the Hamming distance between each annotator's labels and the mode of the annotations as the consistency metric. The specific formula for each movie is defined as follows:

$$D = \frac{1}{10m} \sum_{k=1}^{3} \sum_{i=1}^{m} \sum_{j=1}^{10} \mathbb{I}\left(A_{ij}^{(k)} \neq M_{ij}\right) \tag{1}$$

Here, $A^{(k)}$ denotes the annotation results from the k_{th} annotator, m denotes the number of clips per movie, M represents the mode of the three annotators' labels and $\mathbb{I}(\cdot)$ denotes the indicator function.

Since binary classification is used, for each position in the annotation matrix, there will always be at least two annotators who agree, which may bias the consistency results. To address this, we first compute the sum of the Hamming distances between each annotator's labels and the mode, and then normalize the result. This ensures that the final consistency score falls within the range $[0, 1]$, where values closer to 0 indicate higher consistency among the three annotators. Finally, we average the Hamming distances over all ten emotions across all movies, resulting in a single value that reflects the overall inconsistency of the dataset, which is 0.41.

To further investigate the potential causes of annotation inconsistency, we analyzed the consistency among annotators of the same gender. Since each movie is annotated by three annotators,

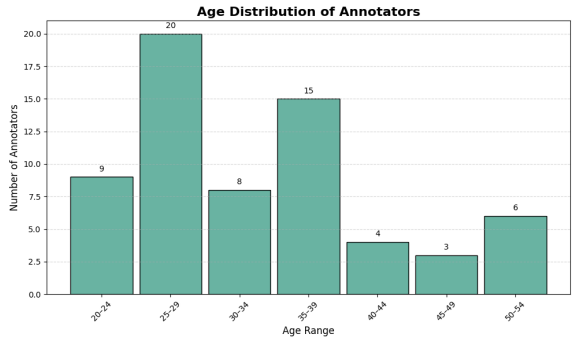

Figure 1: Age distribution of annotators.

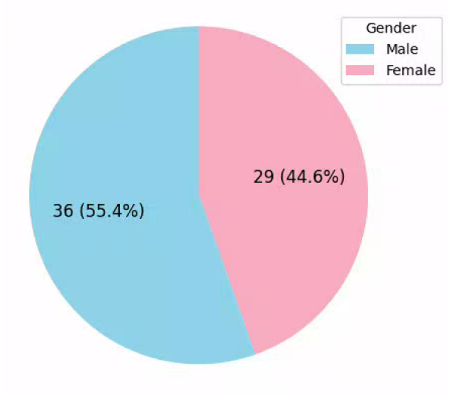

Figure 2: Gender distribution of annotators.

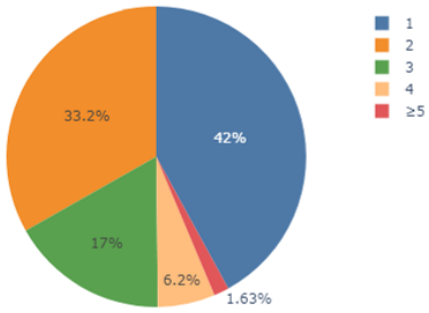

Figure 3: The proportion of video clips annotated with different numbers of emotion labels.

there are always at least two annotators of the same gender. If all three annotators share the same gender, the consistency metric for that movie is calculated using the same method as (1). If only two annotators are of the same gender, we compute the normalized Hamming distance between their annotations to represent the consistency for that movie. Finally, we average the consistency scores across all movies. The results show that the inconsistency among male annotators is 0.25, while that among female annotators is 0.22. These findings suggest that gender differences can influence how annotators perceive and interpret emotional content in videos—annotators of the same gender tend to reach higher agreement.

In addition, we further investigated the consistency of same-gender annotators with respect to each specific emotion label. For each movie, if exactly two of the three annotators share the same gender, the consistency score is calculated using the following formula:

$$D_j = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\left(A_{ij}^{(1)} \neq A_{ij}^{(2)}\right) \tag{2}$$

Here, $A^{(k)}$ denotes the annotation results from the k_{th} annotator, m denotes the number of clips per movie and $\mathbb{I}(\cdot)$ denotes the indicator function.

To compute the inconsistency score for a specific emotion, we extract the corresponding column from the annotation matrix and calculate the Hamming distance based solely on that dimension. If all three annotators are of the same gender, we compute the Hamming distance between each annotator's labels and the mode result, sum these distances, and normalize the total. By averaging the consistency scores across all movies, we obtain the results shown in Table 1.

Although female annotators generally exhibit lower overall inconsistency, their inconsistency is notably higher when labeling trust. This aligns with findings from existing psychological research, which indicate that women tend to show greater variability in interpreting trust-related emotions [7]. In addition, previous studies have shown that the perception of surprise can be affected by the interaction between age and gender, which may lead to instability in its recognition [1]. This is consistent with our experimental results, where both male and female annotators exhibited relatively high inconsistency when evaluating the emotion of surprise. On the other hand, the high level of agreement among female annotators in identifying positive emotions in videos is also in line with prior psychological research findings [2, 4].

In addition, we also evaluated the consistency of annotations for Valence and Arousal. Given that each of the two labels has three possible annotation outcomes, we evaluate annotation consistency by computing the accuracy between each annotator and the majority label. The final average consistency scores across all movies were 0.8738 for Valence and 0.7817 for Arousal. We also analyzed the consistency of Valence and Arousal annotations by annotator gender, and the results are presented in Table 2. The statistical analysis suggests that there is no significant difference between male and female annotators in terms of their consistency when labeling Valence and Arousal. Interestingly, in some cases, the consistency between annotators of different genders was even slightly higher than that between same-gender annotators. This finding indicates a potential gender-neutral effect in the perception and annotation of Valence and Arousal.

Finally, although we collected the ages of the annotators, there are many possible ways to categorize age groups. Moreover, for each movie, the three annotators do not necessarily include two or more individuals within the same age group. This significantly reduces the number of movies that meet the criteria for age-based analysis, thereby diminishing the statistical validity and value of such

Unpublished working draft. Not for distribution.

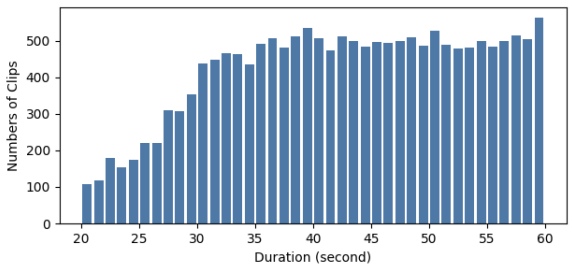

Figure 4: The duration distribution of clips.

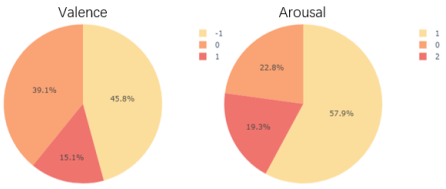

Figure 5: The statistical results of Valence and Arousal annotations in the dataset.

experiments. Therefore, we only performed a statistical analysis of the age distribution, without further exploring the relationship between annotators' age and their labeling behavior, as shown in Figure 1.

B Introduction to the Three Baseline Models

AMP enhances representation learning through two adversarial strategies: adversarial temporal masking and adversarial parameter perturbation. The former improves cross-modal encoding by masking temporal segments adversarially, while the latter strengthens generalization by applying adversarial perturbations to model parameters.

CARAT introduces a reconstruction-based fusion mechanism to better capture fine-grained modality-to-label dependencies. By contrastively learning modality-separated and label-specific features, CARAT enhances multimodal representation. Additionally, a novel sample-wise and modality-wise shuffle strategy is applied to the reconstructed embeddings to enrich label co-occurrence dependencies.

TFN models both intra-modality and inter-modality dynamics in an end-to-end manner. Inter-modality interactions are captured via a tensor fusion approach that explicitly aggregates uni-modal, bimodal, and tri-modal features. Intra-modality dynamics are modeled using three modality-specific embedding subnetworks for language, visual, and acoustic inputs, respectively.

C Experiment Setup

C.1 Definition

Let $X^m \in F^{t_m \times d_m}$ represents the feature of a modality. m can represent audio (a), visual (v) and text (t) in the dataset. t_m represents the temporal length of a modality, and d_m represents the feature dimension of a modality. Let $Y = \{y_1, y_2, \ldots y_p\}$ represent the label space with p labels. For experiments about arousal and valence, p is equal to 3, while p is equal to 2 for the experiments about primary emotion. The task of classification is to predict the corresponding label y_i given the input features of i_{th} clip X_i^m, where $m \in \{a, v, t\}$.

C.2 Dataset Split

The training, validation, and test sets are split with a ratio of 0.7, 0.1, and 0.2. We partition randomly our dataset five times and perform five replicate trials to enhance confidence for experimental results. All clips from the same movie in each division will only appear in one of the training, validation and test sets at the same time. Each split of the training set, validation set, and test set contains all label categories, with the label distribution proportions being roughly the same.

C.3 Feature Extraction

Video clips in our dataset express emotions in a multimodal manner. Therefore, we separately extract visual, auditory, and text modality features for video affective content analysis. For the visual modality, each video is uniformly sampled at 8 frame increments to reduce the computational cost. Then, we extract 768-dimensional visual features using CLIP-large [6], which can capture universal representations. In the case of the audio modality, FFMpeg is employed to extract audio from the original movie video, with the audio format subsequently unified to 16 kHz. Subsequently, VGGish [5] is used to extract 128-dimensional audio features. With regards to the text modality, Tencent Cloud Automatic Speech Recognition (ASR) is employed to extract subtitles. Subsequently, the Chinese-MacBERT-large [3] model is employed to extract 1024-dimensional text features.

C.4 Implementation Details

The specific parameter settings for each baseline are as follows. For AMP, in the PE experiment, the learning rate is set to 1e-5, weight decay to 1e-2, training for 30 epochs, and using AsymmetricLoss as the classification loss function. In the VA experiment, the learning rate is set to 5e-5, weight decay to 1e-2, training for 20 epochs, and using weighted BCE Loss. In both cases, the batch size is 16, and the hidden feature dimension is 256. For CARAT, in the PE experiment, using AsymmetricLoss as the classification loss function. In the VA experiment, using weighted BCE Loss. In both cases, the batch size is 64, hidden feature dimension is 256, learning rate is set to 5e-5, weight decay to 1e-2 and training for 20 epoch.

For TFN, the learning rate is set to 1e-3, training for 30 epochs, weight decay to 1e-5. We use MSEloss as the classification loss function. The batch size is 32. All models were trained on NVIDIA GeForce RTX 3090.

For every emotion,we will get a regression value.And we set thresholds to classify.In the PE experiment,the threshold is set to [0.25,0.3,0.2,0.2,0.3,0.08,0.12,0.1,0.07,0.07].In the VA experiment, the threshold for V is set [-0.6,0.1] and the threshold for A is set [0.8,1.3].

D Comparison with Text Generated by Large Language Models

During the dataset construction process, we observed that a number of recent studies have begun to use large language models

Unpublished working draft. Not for distribution.

Table 1: Inter-Annotator Consistency Analysis of Ten Basic Emotion Labels: Experimental Results (Lower Values Indicate Higher Consistency)

Annotators' genders	Primary Emotion Labels										
	Anger	Surprise	Fear	Joy	Sadness	Shame	Disgust	Trust	Happiness	Amusement	All emotions
Male	0.25	0.31	0.25	0.24	0.25	0.23	0.26	0.28	0.22	0.24	0.25
Female	0.20	0.37	0.25	0.18	0.20	0.20	0.24	0.28	0.13	0.17	0.22
All sexes	0.40	0.58	0.44	0.38	0.39	0.38	0.42	0.47	0.32	0.36	0.41

Table 2: Inter-Annotator Consistency Analysis of Ten Basic Emotion Labels: Experimental Results (Higher Values Indicate Higher Consistency)

Annotators' genders	Valence	Arousal
Male	0.7468	0.6280
Female	0.8366	0.6364
All sexes	0.8738	0.7817

(LLMs) to assist in dataset creation. Inspired by this line of work, we also attempted to leverage an LLM to further expand the scale of our dataset. Specifically, we selected a state-of-the-art large language model, Qwen3-Omni, and conducted experiments using its generated movie content descriptions.

We provided the model with the application context of the text modality in our dataset, instructing it to generate movie descriptions that could be understood by visually impaired audiences. The specific prompt is provided in the Appendix D. The prompt is shown as follows.

Prompt Provided to Qwen3-Omni

You are a professional film narration scriptwriter. Please watch the video content and generate a complete and coherent Chinese textual description for visually impaired audiences.

Writing requirements:

(1) Describe the visual content using natural and conversational Chinese;
(2) Narrate the main actions, character expressions, and environmental changes in chronological order;
(3) Avoid repetitive or redundant descriptions, and do not repeat the same sentences;
(4) Do not output any analysis, numbering, or explanations; only provide the complete textual description;
(5) The description should be sufficiently detailed to enable visually impaired audiences to clearly imagine the entire scene.

The newly generated descriptions were then used to replace the original text modality in our dataset, and experiments were conducted using three baseline emotion analysis models. The final experimental results are shown in Table 3.

Table 3: Experimental results of using text generated by Qwen3-Omni on three models. The values in bold indicate that, in the corresponding experiment, the large language model–generated text outperforms the human-annotated text.

Label	ACC			F1 or weighted-F1		
	AMP	CARAT	TFN	AMP	CARAT	TFN
Primary Emotion	0.2566	0.2559	0.2501	0.4320	0.4327	0.4199
Valence	0.6066	0.6224	0.5668	0.6106	0.6203	0.5684
Arousal	0.5182	0.6041	**0.5683**	0.5147	**0.6084**	**0.5764**

The results indicate that, although the LLM-generated texts exhibit certain advantages in terms of the arousal, their overall performance is inferior to that achieved using human-annotated texts. By examining the generated outputs in detail, we found that the LLM-generated descriptions occasionally contain repetitive content. Moreover, for some video segments, the model failed to generate corresponding textual descriptions regardless of the number of attempts. Consequently, we decided to abandon the direct use of LLM-generated movie narration for dataset expansion. In future work, with the inclusion of a human review process, we plan to explore the use of large language models to generate movie narrations in multiple languages, thereby further expanding the dataset.

blished working draft. for distribution

References

[1] Vladimir A Barabanschikov and Ekaterina V Suvorova. 2021. Gender differences in the recognition of emotional states. *ÏÑÑÒÍÊÎÂÅ× ÅÑÊÀß ÍÀÓÊÀ È ÍÀÐÀÇÎÂÀÍÈÅ PSYCHOLOGICAL SCIENCE AND EDUCATION* 26, 6 (2021), 108.

[2] Olivier Collignon, Simon Girard, Frederic Gosselin, Dave Saint-Amour, Franco Lepore, and Maryse Lassonde. 2010. Women process multisensory emotion expressions more efficiently than men. *Neuropsychologia* 48, 1 (2010), 220–225.

[3] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. doi:10.18653/v1/2020.findings-emnlp.58

[4] Uta-Susan Donges, Anette Kersting, and Thomas Suslow. 2012. Women's greater ability to perceive happy facial emotion automatically: gender differences in affective priming. *PloS one* 7, 7 (2012), e41745.

[5] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi:10.1109/icassp.2017.7952132

[6] Alec Radford, JongWook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Askell Amanda, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *Cornell University - arXiv,Cornell University - arXiv* (Feb 2021).

[7] Yan Wu, Alisha SM Hall, Sebastian Siehl, Jordan Grafman, and Frank Krueger. 2020. Neural signatures of gender differences in interpersonal trust. *Frontiers in human neuroscience* 14 (2020), 225.