Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36929

Research on Public Cognition Prediction Models for Audiovisual CSR Communication Based on Multimodal Semantic Fusion

Yiwen Zhong

School of Arts and Cultures, Newcastle University, Newcastle, UK
wellington589125@gmail.com

Abstract. Digital platforms have transformed CSR communication from textual disclosure into audiovisual and interactive expression, where public cognition is shaped by visual, textual, audio, and platform features. This study combines film-based audiovisual analysis with multimodal learning to predict public cognition of CSR videos from 2020–2024. Key frames, subtitles, comments, audio affect, and interaction data were extracted, while cognition labels covered responsibility credibility, emotional empathy, and action intention. The cross-modal Transformer achieved 0.861 Accuracy, 0.858 Macro F1, and 0.902 AUC, outperforming single-modality and traditional fusion methods. Textual, visual, audio, and interaction features contributed 34.6%, 31.2%, 18.5%, and 15.7%, respectively. The findings show that public cognition results from the combined effects of responsibility narration, visual evidence, sound emotion, and platform interaction, providing a computational approach to audiovisual interpretation in film studies.

Keywords: Audiovisual CSR Communication, Multimodal Semantic Fusion, Public Cognition Prediction, Film Studies, Machine Learning

1. Introduction

Digital platforms have shifted corporate social responsibility communication from written disclosure to audiovisual, emotional, and narrative expression. Public service videos, branded documentaries, and corporate image films have become important media through which audiences understand corporate ethics. From the perspective of film studies, camera language, sound design, character relations, and narrative structure shape the production of meaning, while CSR communication research focuses on how responsibility information affects public trust, attitudes, and behavioral intention [1]. Existing studies have mainly examined CSR videos through communication effects, brand reputation, or stakeholder management, leaving the interaction between audiovisual signs and public cognition insufficiently integrated [2]. Developments in computer vision, natural language processing, and audio affect analysis make it possible to transform audiovisual content into computable multimodal semantic features. In this context, public videos, subtitles, audio features, audience comments, and survey evaluations can be combined to construct a multimodal semantic

DOI: 10.54254/2755-2721/2026.36929

fusion model for analyzing how different audiovisual elements influence responsibility credibility, emotional identification, and action intention.

2. Literature review

2.1. Audiovisual CSR communication and public cognition

The core of CSR communication lies not only in the external release of responsibility information, but also in how audiences form judgments about corporate motivation, value orientation, and the authenticity of social action during media exposure. According to previous studies, the communication effectiveness of the responsibility largely depends on such aspects as information clarity, issue relevance, and the consistency of the claims and behavior of the corporation, where audiences judge the level of trust in terms of the comparison of the communicated content and the real one [3]. Within the context of audiovisual communication, CSR information becomes a part of the human stories, social settings, and emotions, so the judgments about the credibility of responsibility become the result of the influence of all three factors together. The public service advertising and documentaries enhance the role of storytelling for increasing the involvement of the audience making the responsibility claims perceptible within the social experience [4].

2.2. Film semiotics and audiovisual emotional expression

The semiotics of film studies the ways in which composition, movement of the camera, tempo of editing, color relations, and sound construction affect audience perception and produce meaning through the audiovisual organization. Often CSR videos create authenticity using documentary visuals, lower saturation of colors, natural light, and close-ups of the characters while music, narration, and environmental sounds help to set the emotional tone of responsibility issues [5]. The way beneficiaries, community spaces, and corporate agents are represented affects the way audiences perceive the responsibility relations and form their moral perception of corporations' actions. Multimodal discourse analysis reveals the social meaning as constructed by the collaboration of visual grammar and language narrations and proves that images are independent semiotic expressions of value and not just visual complements of the text [6].

2.3. Multimodal semantic fusion and cognition prediction

Multimodal learning deals with the modeling of visual, textual, and audio data at the same time, thus being an appropriate tool for the analysis of CSR videos containing visual storytelling, subtitles, speech, and musical emotion. Visual models are capable of detecting scenes, people, and action signals; language models are capable of extracting responsibility themes, value claims, and attitudes toward comments; while audio characteristics might represent emotional strength, rhythmic variations, and storytelling atmosphere [7]. In case the semantically complementary features from different modalities are found in the same communication object, then the fusion models would better capture public cognition variations than unimodal models. Cross-modal Transformer and Attention networks allow estimating the contribution of each modality to the prediction task, thus helping us understand what audiovisual features are responsible for credibility, identification, and action intention [8]. The combination of model results with film-semiotic analysis will help to see the connection between algorithmic features and audiovisual meaning [9].

DOI: 10.54254/2755-2721/2026.36929

3. Experimental methods

3.1. Data collection and label construction

Data were collected from corporate CSR sections, Bilibili, Douyin, Weibo Video Account, and public CSR case libraries between 2020 and 2024, covering philanthropy, environmental responsibility, rural revitalization, education, and community service. Of 520 videos initially collected, 480 valid samples remained after excluding incomplete, low-resolution, low-comment, or duplicate content. Recorded variables included video metadata, subtitles, audio, keyframes, engagement metrics, and comments.

Public cognition labels combined questionnaire ratings from 120 participants on responsibility credibility, emotional identification, and action intention with manually verified comment sentiment and topic consistency. Two coders independently reviewed the labels, achieving a Cohen's Kappa of 0.82, indicating sufficient reliability for model training.

3.2. Multimodal feature extraction and semantic fusion

Multimodal feature extraction was conducted through four paths, including visual data, textual data, audio data, and comment feedback. For the visual modality, keyframe sequences were generated by extracting one frame every 5 seconds. CLIP was used to extract scene, character, and action semantic vectors. For the textual modality, subtitles, titles, and comments were processed through word segmentation, noise removal, and stop word filtering. BERT was then used to generate textual semantic vectors [10]. For the audio modality, audio data were converted into 16 kHz mono format. MFCC, loudness, rhythm, and fundamental frequency features were extracted to identify emotional intensity and sound rhythm. Different modalities were standardized before entering the joint modeling stage. The single modality feature representation is shown in Formula 1.

$$h_i = f(W_i x_i + b_i), \quad i \in \{v,t,a\} \tag{1}$$

In Formula 1, x_i denotes the input feature of the i_{th} modality. v, t, and a denote the visual, textual, and audio modalities, respectively. W_i denotes the modality mapping matrix. b_i denotes the bias term. $f(\cdot)$ denotes the nonlinear activation function. h_i denotes the hidden semantic representation of the iiith modality. In the fusion stage, a cross modal attention mechanism was used to calculate modality weights. This allowed the model to dynamically judge the importance of visual evidence, textual responsibility themes, and sound emotion according to the expression structure of different CSR videos. The joint semantic representation is shown in Formula 2.

$$h_f = \sum_{i \in \{v,t,a\}} \alpha_i h_i, \quad \alpha_i = \frac{\exp(q^\top h_i)}{\sum_{j \in \{v,t,a\}} \exp(q^\top h_j)} \tag{2}$$

In Formula 2, h_f denotes the fused multimodal semantic vector. α_i denotes the attention weight of the i_{th} modality. q denotes the trainable query vector. h_i and h_j denote the hidden representations of different modalities. This process was used to measure the relative contribution of different audiovisual modalities in public cognition prediction.

DOI: 10.54254/2755-2721/2026.36929

3.3. Model training and evaluation design

Model training used the 480 valid videos described above as the basic units for dataset division. The training set, validation set, and test set were divided in a ratio of 70 percent, 15 percent, and 15 percent, corresponding to 336, 72, and 72 videos. The division process used each video as the smallest unit. This prevented different clips from the same video from entering the training set and test set at the same time, reducing data leakage. Two tasks were designed. The first was a classification task for high and ordinary public cognition. Samples with an average score above 3.80 across responsibility credibility, emotional identification, and action intention were labeled as the high cognition group. The remaining samples were labeled as the ordinary cognition group. The second was a regression task for public cognition scores, which directly predicted the average score of the three dimensions. During training, the AdamW optimizer was used. The learning rate was set at 2×10^{-5}, the batch size was set at 16, and the maximum number of training epochs was set at 30. Early stopping was applied when the validation loss did not decrease for 5 consecutive epochs. Accuracy, Macro F1, and AUC were used for the classification task. MAE, RMSE, and R^2 were used for the regression task [11]. The classification prediction function is shown in Formula 3.

$$p(y=1|h_f)=\sigma(W_o h_f+b_o) \tag{3}$$

In Formula 3, $p(y=1|h_f)$ denotes the probability that a sample is predicted as the high public cognition category. h_f denotes the fused multimodal semantic vector. W_o denotes the output layer weight matrix. b_o denotes the output layer bias term. $\sigma(\cdot)$ denotes the Sigmoid function. y denotes the public cognition category label.

4. Results

4.1. Prediction performance comparison

Under the same train-validation-test split, visual, textual, audio, early fusion, attention fusion, and cross-modal Transformer models were evaluated on identical samples. The textual model achieved an F1 of 0.782, showing strong performance in identifying CSR themes and public attitudes. The visual and audio models reached F1 scores of 0.744 and 0.701, respectively, contributing scene-action and emotional information.

As shown in Table 1, the cross-modal Transformer achieved 0.861 Accuracy, 0.858 F1, and 0.902 AUC, with RMSE reduced to 0.416, outperforming both fusion baselines. The results indicate that public cognition is jointly shaped by visual evidence, responsibility narration, and sound emotion, and that multimodal modeling better captures their semantic complementarity.

Table 1. Comparison of public cognition prediction performance across different models

Model	Accuracy	Macro F1	AUC	MAE	RMSE	R^2
Visual only CLIP	0.751	0.744	0.803	0.472	0.586	0.412
Text only BERT	0.789	0.782	0.842	0.438	0.531	0.486
Audio only MFCC	0.713	0.701	0.764	0.506	0.628	0.351
Early Fusion	0.816	0.811	0.871	0.397	0.482	0.562
Attention Fusion	0.839	0.835	0.889	0.372	0.447	0.604

DOI: 10.54254/2755-2721/2026.36929

Table 1. (continued)

Cross modal Transformer	0.861	0.858	0.902	0.348	0.416	0.653

4.2. Interpretation of multimodal contribution

To further explain the role of different modalities in public cognition prediction, attention weights and mean SHAP values were analyzed on the test samples. The textual modality showed the highest average contribution at 34.6%, mainly from responsibility themes, value judgments, and public attitudes in subtitles, titles, and comments. The visual modality contributed 31.2%, reflecting visual evidence such as real characters, community scenes, corporate actions, and beneficiaries. The audio modality contributed 18.5%, mainly supporting the judgment of emotional intensity, while interaction indicators contributed 15.7% through likes, shares, and comment density. As shown in Figure 1, the bubble matrix indicates that text contributed most to responsibility credibility, reaching 36.8%, while visual features contributed 32.1% to emotional identification. Interaction indicators had a higher contribution to action intention, reaching 18.9%. These results suggest that public cognition is jointly shaped by responsibility narration, visual evidence, sound emotion, and platform feedback. High cognition samples usually contained clear action presentation, strong character relations, and moderate emotional expression, whereas low cognition samples often showed slogan like subtitles, excessive corporate self presentation, and insufficient visual evidence of responsibility actions.

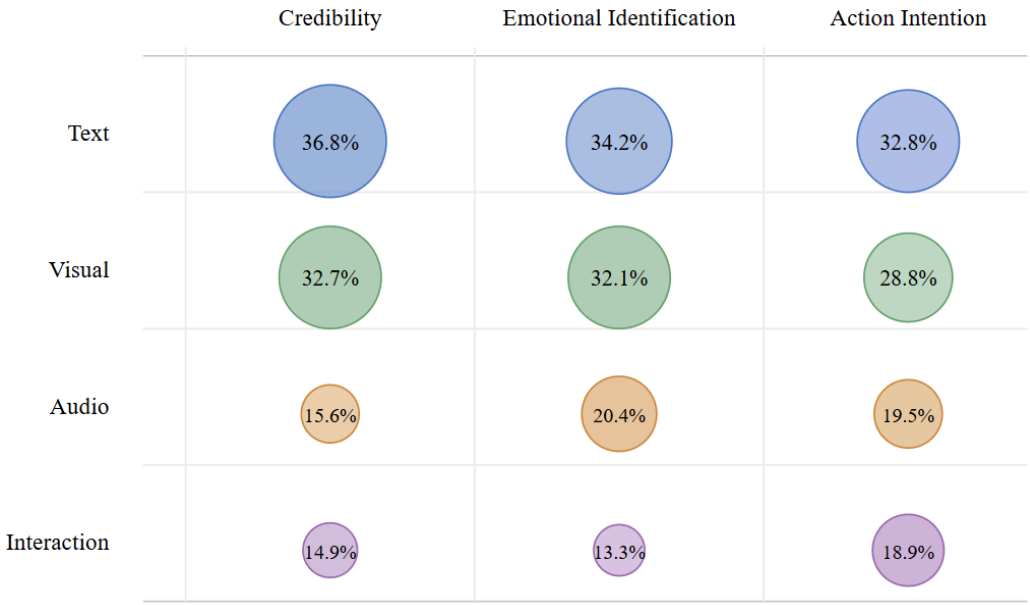

Figure 1. Bubble matrix of multimodal contribution to public cognition

5. Discussion

The experimental results show that multimodal semantic fusion can effectively explain differences in public cognition in audiovisual CSR communication. The textual modality carried a higher weight in identifying responsibility themes and value judgments. The visual modality enhanced responsibility credibility through characters, scenes, and action evidence. The audio modality mainly supplemented the judgment of emotional intensity. This result is consistent with the film studies

DOI: 10.54254/2755-2721/2026.36929

view that shots, sound, and narrative jointly produce meaning. Platform interaction indicators also contributed to action intention, indicating that public cognition is affected by the communication environment and social feedback. Although the multimodal model improved prediction accuracy, the interpretation of results still needs to return to audiovisual narrative, corporate ethics, and audience reception context, so that complex viewing experience is not reduced to technical indicators.

6. Conclusion

Public cognition prediction in audiovisual CSR communication can be organized through a complete technical process, including public video collection, label construction, feature extraction, semantic fusion, and model interpretation. The results show that the cross-modal Transformer performed well in both classification and regression tasks, indicating stable semantic complementarity among visual, textual, audio, and interaction data. The framework connects audiovisual semiotic analysis in film studies with computer-based multimodal modeling. It can identify key narrative factors in CSR videos and quantify their influence on public cognition. Future research can expand enterprise categories, platform sources, and cross-cultural samples. Long-term behavioral data can also be introduced to improve model generalization and explanatory power in communication analysis.

References

[1] Zhao, W., Cheng, Y., & Lee, Y.-I. (2023). Exploring 360-degree virtual reality videos for CSR communication: An integrated model of perceived control, telepresence, and consumer behavioral intentions. *Computers in Human Behavior, 144*, 107736.

[2] Macca, L. S., et al. (2024). Consumer engagement through corporate social responsibility communication on social media: Evidence from Facebook and Instagram Bank Accounts. *Journal of Business Research, 172*, 114433.

[3] Inversini, A., & Derchi, G. B. (2024). Corporate social responsibility on social media: A scoping review of the literature. *Journal of Information, Communication and Ethics in Society, 22*(4), 434-452.

[4] Dong, C., et al. (2024). Constructing care-based corporate social responsibility (CSR) communication during the COVID-19 pandemic: A comparison of Fortune 500 companies in China and the United States. *Journal of Business Ethics, 192*(4), 775-802.

[5] Nguyen, N., et al. (2023). CSR-related consumer scepticism: A review of the literature and future research directions. *Journal of Business Research, 169*, 114294.

[6] Cao, P., et al. (2024). Eco-engagement: Tracing CSR communication's ripple effect on consumer hospitality loyalty. *Journal of Retailing and Consumer Services, 79*, 103879.

[7] Zhou, J., et al. (2024). The role of social media in CSR performance: An integrated institutional and resource dependence perspective. *Journal of Business Research, 184*, 114880.

[8] Dong, H., Zhao, P., & Ngai, C. S. B. (2025). Engaging social media users with corporate social responsibility messages: An integrated theoretical approach in the automotive industry during social media era. *PLOS ONE, 20*(6), e0322481.

[9] Erokhin, D. (2025). ESG reporting in the digital era: Unveiling public sentiment and engagement on YouTube. *Sustainability, 17*(15), 7039.

[10] Illia, L., Ballester-Ripoll, R., & Clausen, A. K. (2025). Fabricating CSR authenticity: The illusory truth effect of CSR communication on social media in the AI era. *Public Relations Review, 51*(3), 102588.

[11] Kim, Y., Acaf, Y., & Kim, J. (2025). Fast fashion's environmental CSR through NGO collaboration: The impact of collaboration level and environmental focus on CSR authenticity and consumer responses. *Public Relations Review, 51*(3), 102585.