# Exploring Visual, Textual, And Engagement Features for Memorability Predictions

Mahnoor Adeel[1,*,†], Kisa Fatima[2,†], M. Ibrahim Ayoubi[3,†], Mustafa Usmani[4,†] and Muhammad Atif Tahir[5,†]

[1]School of Mathematics and Computer Science, Institute of Business Administration (IBA), Karachi, Pakistan

Abstract

This project explores the task of predicting video and advertisement memorability, including brand memorability based on a specific ad, using machine learning techniques. The study involved analyzing multimodal data comprising textual, visual, and numerical features. Unlike much of the existing work that focuses on single modalities, this study emphasizes the integration of multimodal representations through both feature-level and prediction-level fusion. Key contributions include the frame-wise visual regression with statistical fusion, the use of contextual transformer embeddings for textual fields, incorporation of engagement-based numerical features, and an optimal weighted ensemble strategy to combine modalities. These innovations collectively improve prediction accuracy and demonstrate the value of combining diverse feature types and regression models in ad memorability prediction.

## 1. Introduction

Memorability plays a key role in determining how effectively visual and narrative content resonates with an audience. In advertising and media, understanding what makes certain content more memorable can provide valuable insights for improving audience engagement, brand recall, and creative impact.

This paper presents the participation of the **CVG-IBA team** in the MediaEval 2025 Memorability Challenge [1], which focuses on predicting memorability across two major tasks involving both videos and advertisements.

The video-related tasks involved predicting overall video memorability and EEG responses reflecting viewer engagement, while the advertisement-related tasks focused on predicting ad and brand memorability. The performance of all models was measured using Mean Squared Error (MSE) to evaluate prediction accuracy and Spearman's Rank Correlation Coefficient (SRCC) to assess the consistency of the ranking between predicted and true memorability scores.

For both video and advertisement tasks, a multimodal approach was utilized combining textual descriptions, categorical and numerical metadata, and image features extracted from three representative frames of each video or ad. The work explored how multimodal representations and fusion methods can effectively model the factors that make videos and ads memorable.

*MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online*

*Corresponding author.

†These authors contributed equally.

m.adeel.26913@khi.iba.edu.pk (M. Adeel); k.fatima.27076@khi.iba.edu.pk (K. Fatima); m.ayoubi.26269@khi.iba.edu.pk (M. I. Ayoubi); mustafaausmani@gmail.com (M. Usmani); atiftahir@iba.edu.pk (M. A. Tahir)

## 2. Related Work

The understanding of advertisement memorability through multimodal and semantic modeling has seen advancements in recent years. Harini et al. [2] studied long-term ad memorability. They highlighted how visual, textual, and emotional cues influence recall and suggested methods for generating more memorable ads. McCoy and Aka [3] examined how linguistic style and emotional tone in brand slogans affect recall. They found that short and effective phrasing improves memorability.

Asgarian et al. [4] introduced *MindMem*, a multimodal framework combining large language models with deep visual networks for ad memorability prediction. Inspired by this work, our study uses a multimodal approach. We integrate visual, textual, and engagement-based features through optimized fusion and ensemble regression to predict advertisement and brand memorability.

Other relevant works include Multimodal Memorability by Newman et al. [5], which models how memorability decays with time using visual and semantic information, and VideoMem by Cohendet et al. [6], which constructs and analyzes both short-term and long-term video memorability using large-scale datasets.

## 3. Approach

For this task, a multimodal learning framework that utilized both visual and textual information from advertisements was adopted, along with additional numerical engagement metrics. The goal being effective combination of different modalities including images, text, and metadata, to capture the full spectrum of factors influencing ad memorability [7]. The overall pipeline remained similar for different challenges, with minor adjustments depending on the structure of the data and the scope of the prediction.

### 3.1. Visual Features

Each advertisement contained three representative frames, which were treated as visual samples describing the ad's overall tone and composition. For each frame, we extracted deep visual embeddings using pretrained Convolutional Neural Network (CNN) architectures such as VGG, ResNet, and EfficientNet. A separate regressor was trained on the embeddings from each frame to learn frame-specific patterns related to memorability. The outputs from the three regressors were then combined through weighted averaging, with the middle frame given the highest weight as it performed best on the validation dataset.

In addition to simple averaging, statistical fusion techniques were also explored, particularly with ResNet embeddings, which showed strong performance. The fusion combined embeddings using feature-level statistics (mean, standard deviation, min and max) and improved robustness by capturing inter-frame variations. Overall, VGG and ResNet provided consistently good results.

For the regression stage, XGBoost, Ridge Regression, and Random Forest were used. All three performed well, with XGBoost showing strong generalization and Ridge providing stable results on smaller feature sets. Overall, VGG and ResNet produced the most effective embeddings, leading to consistently better performance across regressors.

### 3.2. Textual Features

The textual component of each ad included several fields such as description, channel name, transcription, title, and tags. Multiple representation techniques such as TF-IDF, Word2Vec, and Transformer-based encodings were tested.

Among these, TF-IDF and BERT [8] consistently produced the strongest results, with BERT performing slightly better overall. For transformer-based encodings, all textual fields (e.g., description and transcription) were concatenated using placeholder tokens to maintain structural and contextual boundaries between sections.

Inspired by DeepSORT, we tried an approach for Brand Memorability that reframes brand mention detection in transcripts as a tracking problem rather than simple keyword counting. Each detected mention is considered a "detection," represented by a contextual embedding derived from its surrounding words. These embeddings act as appearance descriptors that capture semantic similarity between mentions, allowing the system to recognize the same brand even when it appears with spelling variations or in slightly different contexts.

Using cosine similarity and the Hungarian algorithm, each new mention is matched to an existing "track" if it is contextually and temporally close; otherwise, a new track is started. Each track represents a consistent stream of brand mentions, updated over time using an exponential moving average to adapt to phrasing shifts. This tracking-based method ensures that mentions are not double-counted or fragmented, producing a robust and context-aware measure of how frequently and consistently a brand is referenced throughout the transcript.

### 3.3. Numerical Features

In addition to image and text data, the dataset included several engagement-based numerical attributes such as durationString, durationSeconds, viewsCount, likesCount, dislikesCount, favouritesCount, commentsCount, engagementsCount, and engagementRate. These features provided useful context on audience interaction and viewing behavior, which often correlate with how memorable an advertisement is. All numerical features were cleaned, converted to numeric form where necessary, and normalized to maintain consistent scaling with other modalities.

### 3.4. Result Aggregation and Optimal Weighting

Final predictions from the visual, textual, and numerical models were combined using optimized weights to minimize MSE. The weights were determined using an objective function, with constraints ensuring they summed to 1 and remained non-negative. The visual modality received the highest weight, indicating its stronger correlation with ad memorability.

### 3.5. Challenge 1.2: Is this person familiar with this video?

Initially, moderate separability was observed while examining the ERP dataset to gain insights into signal characteristics and label distribution. Building on this analysis, ERSP features were incorporated by aggregating frequency-time data per epoch to enhance the ERP information. Various classifiers, including Logistic Regression, Ridge, XGBoost, and AdaBoost, were tested for their ability to generalize across participants. However, some models faced overfitting due to a limited number of positive remembered samples. Addressing this and combating the class imbalance issue, normalization and training of hybrid models using both ERSP and ERP features with XGBoost and Logistic Regression were done while using weighted averages, assigning a higher weight to the model trained on the ERP data. This combined approach

utilized both temporal and spectral EEG dynamics, achieving an AUC of 0.544 on the test set (as mentioned in the results returned), reflecting a balanced tradeoff between model complexity and generalization.

## 4. Results and Analysis

**Table 1:** Best performing runs for each challenge for **CVG-IBA** on test set.

| Task | Textual Embeddings | Visual Embeddings | Models Used | SRCC | MSE/ AUC |
|---|---|---|---|---|---|
| 1.1 | — | EfficientNet and ResNet | XGBoost | 0.332 | 0.063 |
| 1.2 | — | EEG / ERP features | Logistic Regression + XGBoost | — | 0.544 |
| 2.1 | BERT | ResNet50 | AdaBoost + XGBoost | 0.165 | 0.025 |
| 2.2 | BERT | ResNet50 | AdaBoost + XGBoost + Ridge | 0.153 | 0.053 |

Table 1 summarizes the results achieved by our team, **CVG-IBA**. They show that multimodal and ensemble-based strategies generally outperformed single baseline models, though overall correlations remained modest. Visual embeddings, particularly the ones provided through EfficientNet and ResNet architectures, complemented the model's performance the most. Text-based embeddings such as BERT added marginal but consistent improvements when combined through ensemble methods.

The combination of boosting-based models, especially XGBoost and AdaBoost, proved to be the most effective fusion approach, offering stable correlations across both memorability regression and classification tasks. For the Task-1.2 the fusion of the predictions from separate models on ERP and ERSP features stabilised the AUC score. However, despite these improvements, the absolute values of SRCC and AUC remain relatively low, suggesting limitations in feature complementarity and generalization across modalities.

These findings indicate that while multimodal fusion improves consistency and robustness, achieving high predictive accuracy remains challenging. The moderate correlations and MSE values reveal that current approaches capture only partial aspects of memorability, and future work should focus on deeper cross-modal alignment and more expressive feature representations.

## 5. Conclusion

This study demonstrates that multimodal fusion of visual, textual, and engagement features enhances memorability prediction. Future work should explore deeper multimodal integration and representation learning to achieve stronger generalization and interpretability.

## 6. Declaration on Generative AI

During the preparation of this work the authors used ChatGPT and Grammarly for grammar checking, paraphrasing, and formatting. All research ideas, experimental work, and interpretations were solely carried out and verified by the authors. They reviewed and edited all content and take full responsibility for what is published.

## References


[1] I. Martin-Fernandez, M. G. Constantin, C.-H. Demarty, M. Gil-Martin, S. Halder, B. Ionescu, A. Matran-Fernandez, R. Savran Kiziltepe, A. Garcia Seco de Herrera, Overview of the mediaeval 2025 predicting movie and commercial memorability task, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.

[2] S. I. Harini, S. Singh, Y. K. Singla, A. Bhattacharyya, V. Baths, C. Chen, R. R. Shah, B. Krishnamurthy, Long-term ad memorability: Understanding & generating memorable ads, arXiv preprint arXiv:2309.00378 (2023). URL: https://arxiv.org/abs/2309.00378, submitted 1 September 2023; current version v5 (30 November 2024).

[3] J. McCoy, A. Aka, Predicting the memorability of brand slogans, SSRN Electronic Journal (2025). doi:10.2139/ssrn.5242034, preprint, January 2025.

[4] S. Asgarian, Q. Jetha, J. Jeon, Mindmem: Multimodal for predicting advertisement memorability using llms and deep learning, arXiv preprint arXiv:2502.18371 (2025). URL: https://arxiv.org/abs/2502.18371.

[5] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, in: ECCV, 2020. URL: https://arxiv.org/abs/2009.02568.

[6] R. Cohendet, C.-H. Demarty, N. Q. K. Duong, M. Engilberge, Videomem: Constructing, analyzing, predicting short-term and long-term video memorability, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2531–2540. URL: https://www.emergentmind.com/papers/1812.01973.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning (ICML 2021), volume 139, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805.