Multimodal Learning Framework for Commercial Video Memorability Prediction

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

In this work, we outline our approach for predicting memorability for videos and advertisements, which is a part of the MediaEval 2026 challenge by the CVG-IBA team. In this study, we examine various ways of incorporating multimodal representations, including visual, textual, color palette, and numeric engagement features using a combination of feature-level and prediction-level fusion techniques. Contrary to approaches based on only one modality, we employ visual regression at a per-frame level alongside statistical and contextual fusion methods as well as the optimal weighting of an ensemble of prediction methods. New features for the color palette obtained via the Binned-KMeans technique have been additionally incorporated in Challenge 1.1.

1. Introduction

Memorability is a key cognitive attribute which decides the degree to which visual and narrative material affects and sticks with the audience. With respect to advertisements and the media industry, knowledge about the attributes that affect memorability can prove to be extremely useful.

In this paper, we describe the contribution made by the **CVG-IBA team** to the MediaEval 2026 Memorability Challenge [1]. Continuing our previous success in the MediaEval 2025 version of the challenge, in which our team obtained the first place both in Subtask 1.1 and Subtask 2.2, we have returned for another year with improved and enhanced approaches. In the current MediaEval edition, our team will participate in two challenges: **Challenge 1.1** and **Challenge 2.1**.

The 2026 MediaEval introduces optional new research directions in Challenge 1.1, such as metadata of movies from the MovieLens 32M corpus, as well as color palettes per clips computed via the Binned-KMeans algorithm [2].

Performance on both the tasks is assessed based on Mean Square Error (MSE), which evaluates prediction error, and Spearman's Rank Correlation Coefficient (SRCC), which measures correlation between the predicted ranks of memorability and actual ranks.

In addressing both problems, a multimodal strategy is chosen that combines the use of texts and metadata along with visual content obtained from representative frames of the videos or commercials.

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, 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

A number of works have considered the prediction of memorability in terms of images, videos, and ads, with the focus of the research shifting from traditional feature engineering approaches towards multimodal deep learning techniques. Constantin et al. [3] conducted a thorough review of the literature on computational memorability, covering both traditional image-level research and recent large-scale multimodal benchmarks, and presenting a unified evaluation framework for measuring progress in memorability research. Isola et al. [4] showed that memorability is an intrinsic property of an image, which has more to do with semantics rather than color and frequency components. Cohendet et al. [5] presented a new benchmark for video memorability analysis, while Newman et al. [6] considered memorability decay modeling through multimodal visual and semantic representations.

Regarding advertising elements, the paper by Harini et al. [7] reported the impact of visuals, text, brands, and audio on long-term ad memorability. Similarly, McCoy and Aka [8] suggested that using clear and emotionally charged phrases contributes to better slogan memorability. Asgarian et al. [9] developed an approach called *MindMem*, where both LLMs and deep visual networks were used to predict ad memorability. Pramov [10] presented a Gemma-3-based fusion approach with LoRA model adaptation from the rationale prompts provided by experts, which achieved remarkable results in the MediaEval 2025 competition. Furthermore, Martín-Fernández et al. [11] demonstrated that parameter-efficient adaptation of large vision-language models via QLoRA yields competitive memorability prediction, highlighting the growing role of LVLMs in this domain.

Color is an emerging yet underexplored feature for memorability. Even though low-level color statistics are not strong enough predictors [4], structured color representations contain meaningful information. The *Binned-KMeans* framework proposed in [2] allows extracting explainable weighted color palettes from videos, which we leverage as an additional modality for Challenge 1.1 of MediaEval 2026.

In this paper, we continue along the aforementioned lines by considering a combination of visual, text, meta information, and color palette features via ensemble regression modeling.

3. Approach

A multimodal learning paradigm was employed in both problems to account for a range of aspects that affect video and ad memorability [12]. While the general approach was similar in both cases, some changes were made to adapt to the specificities of the data and predictions involved.

3.1. Visual Features

Each video was represented by three representative frames, serving as visual representations of the tone, structure, and content of the video. These frames had deep visual embeddings computed based on CNN models such as VGG, ResNet, and EfficientNet, which perform exceptionally well on visual tasks and can be effectively used for extracting features related to memorability. For each of the frames, separate models for predicting frame-specific memorability were trained. The predictions made by these three models were then aggregated by taking the weighted average.

Apart from averaging, several statistical fusions approaches were tested especially when it came to ResNet features. This strategy combined inter-frame data by making use of descriptive statistics including average, standard deviation, minimum and maximum values in order to

achieve better and more complex representation of visual features. The algorithms used in the regression step include XGBoost, Ridge regression and Random Forest regressor. XGBoost showed its good generalization capabilities for any set of features used, and the Ridge Regression method was quite reliable when it comes to small or sparse feature space. In both cases, VGG and ResNet performed the best regarding visual features encoding and prediction accuracy.

3.2. Textual Features

Several text features were associated with each video, including description, channel name, transcription, title, and tags, where all features were treated as possible sources of semantic data for memorability. Several methods of textual representation were examined, among which TF-IDF, Word2Vec, and transformer encoding showed the highest informativity.

In particular, TF-IDF and BERT [13] proved to be more efficient. The latter method demonstrated better performance as it is capable of identifying deep contextual dependencies in texts. For transforming a text to its vector form using transformer architecture, all textual features were merged into one text, with placeholders to mark the borders between fields.

3.3. Color Palette Features

In order to extract the aesthetic qualities of advertisements through their dominant color palettes, k-means clustering was used to extract K-dominant colors from the frames taken from the sampled videos.

Frame-level descriptors, which were made up of the color cluster centroids generated by the k-means clustering algorithm, were combined into fixed-length video representations based on statistics such as mean, standard deviation, min, max, and median across all frames.

Several regression algorithms were used for modeling the learned visual aesthetics into numerical scores for memorability prediction. Some of these included Support Vector Regression (SVR), Random Forest (RF), Gradient Boosting Regression (GBR), and XGBoost.

3.4. Result Aggregation and Optimal Weighting

The weights were adjusted to reduce the prediction error (MSE) as much as possible, while keeping all weights positive. The weights were then scaled so that they added up to 1 before combining the results from the four methods. In each run, the visual method was assigned the highest weight, highlighting the importance of the method in estimating memorability.

4. Results and Analysis

Tables 1 and 2 summarize the results achieved by our team, CVG-IBA, across both challenges.

Table 1
Challenge 1.1: Movie Clip Memorability Results

Run	Spearman	Pearson	MSE
runeffnetvis-rfpal-lgmmeta	0.225	0.250	0.066
runEFFN-xgbvis-gtrpal-knnmeta	0.153	0.170	0.071
runR3D-xgbvis-svrpal-knnmeta	**0.239**	0.237	**0.066**

For Challenge 1.1, multimodal and ensemble-based strategies yielded modest but positive correlations. Visual features extracted via R3D and EfficientNet architectures contributed most to predictive performance, while the inclusion of color palette features provided complementary

signals, with the choice of palette encoding method — SVR versus GTR — noticeably affecting results. The run combining R3D visual features, SVR palette encoding, and KNN metadata achieved the best Spearman correlation of 0.239 and MSE of 0.066, suggesting that spatiotemporal visual representations paired with structured color information offer the most informative feature combination for movie clip memorability.

Table 2
Challenge 2.1: Commercial Ad Memorability Results

Run	Spearman	Pearson	MSE
runBERTXGBoostResNet50XGB	-0.043	-0.055	0.105
runMPNetAdaBoostResNetRidgeFramewiseXGB	-0.031	-0.015	**0.103**
runMPNetAdaBoostVGGXGBridgeRandomForestFramewiseXGB	-0.023	-0.064	0.106
runResNetXGBRidgeRandomForestEngagementFeatureCalibration	-0.020	-0.048	0.105
runTFIDFAdaBoostResNetXGBRidgeRandomForestNormalizedLateFusion	**-0.014**	-0.052	0.105

For Challenge 2.1, all submitted runs produced negative Spearman and Pearson correlations, indicating complexity of the challenge of predicting ad memorability. The lowest MSE of 0.103 was achieved by the MPNet-AdaBoost-ResNet-Ridge framewise fusion run, and the least negative Spearman of -0.014 by the normalized late fusion run combining TFIDF, AdaBoost, ResNet, XGBoost, and Ridge regression. The consistently negative correlations across all runs might suggest that the visual, textual, and engagement-based features employed are insufficient to capture the complexity underlying commercial ad memorability.

Overall, these findings indicate that while multimodal fusion improves 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, richer semantic representations, and better generalization strategies, particularly for the advertisement domain.

5. Conclusion

This study demonstrates that multimodal fusion of visual, textual, color, and engagement features enhances memorability prediction. Future work should explore end-to-end multimodal architectures and large vision-language models to achieve stronger generalization and deeper cross-modal understanding.

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. Martín-Fernández, A. Ganesh, M. G. Constantin, C.-H. Demarty, M. Gil-Martín, S. Halder, B. Ionescu, A. Matran-Fernandez, R. Savran Kiziltepe, A. García Seco de Herrera, Overview of the mediaeval 2026 predicting movie and commercial memorability task, in: Proceedings of the MediaEval 2026 Workshop, Amsterdam, The Netherlands and Online, 2026.

[2] A. Pocol, et al., Binned-kmeans: Extracting memorable, explainable, weighted and sorted colour palettes from videos, in: International Conference on Content-Based Multimedia Indexing (CBMI), Dublin, Ireland, 2025.

[3] M. G. Constantin, C.-H. Demarty, C. Fosco, S. Halder, G. Healy, B. Ionescu, S. V. Luncanu, I. Martín-Fernández, A. Matran-Fernandez, R. Savran Kiziltepe, A. F. Smeaton, L.-D. Stefan, L. Sweeney, A. García Seco de Herrera, A review of computational memorability: A benchmark framework, International Journal of Computer Vision 134 (2026) 298. doi:10.1007/s11263-026-02880-6.

[4] P. Isola, J. Xiao, A. Torralba, A. Oliva, What makes an image memorable?, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 145–152. doi:10.1109/CVPR.2011.5995721.

[5] 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.

[6] 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.

[7] 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).

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

[9] 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.

[10] A. Pramov, LLM-based fusion of multi-modal features for commercial memorability prediction, in: Proceedings of the MediaEval 2025 Workshop, 2025. URL: https://arxiv.org/pdf/2510.22829.

[11] I. Martín-Fernández, S. Esteban-Romero, F. Fernández-Martínez, M. Gil-Martín, Parameter-efficient adaptation of large vision–language models for video memorability prediction, Sensors 25 (2025) 1661. doi:10.3390/s25061661.

[12] 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.

[13] 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.