Unified Modeling of Video and Brand Memorability Using Multimodal and EEG Features

J Bhuvana[1], Ramanan Mahendran[1], S Siddharth Chandrasekar[1], J Pragatheesh[1]

[1] *Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, India*

Abstract

This paper presents a unified approach to predicting video and commercial memorability across four MediaEval 2025 subtasks: movie memorability, EEG-based recall detection, multimodal memorability prediction, and brand memorability. For visual memorability, features from ViT and R3D encoders are combined and used to train stacked regressors combining SVR, XGBoost, and Gradient Boosting, achieving robust performance through feature selection and cross-validation. For EEG-based recall ERP and ERSP features are created and Random Forest classifier is trained, demonstrating that neural signals contain predictive cues for recall. For subsequent brand memorability tasks, visual features from AlexNet, VGG, metadata, and Whisper ASR transcripts embedded via SentenceTransformer are integrated, followed by feature selection and stacked regression. The pipeline consistently improves Spearman correlation across tasks, with stacking outperforming unimodal baselines. These results highlight the value of ensemble learning and multimodal fusion for modeling human memory responses to multimedia content.

1 INTRODUCTION

Memorability Memorability is a fundamental property of multimedia content, reflecting how likely a video or commercial is to be remembered by viewers after exposure. Understanding and predicting memorability have broad applications in advertising, film editing, recommendation systems, and cognitive media research. The MediaEval 2025 Predicting Movie and Commercial Memorability [1] task provides a benchmark for evaluating computational approaches to this problem, encouraging reproducible experimentation across diverse modalities.

This work focuses on predicting both movie scene and commercial memorability using models trained on features extracted from video frames and segments. The approach leverages deep visual encoders such as ResNet, VGG, and Vision Transformers (ViT) to capture semantic and structural cues from scenes. By combining these feature sets, this work aims to exploit the strengths of both convolutional and transformer-based architectures.

To implement this, a pipeline was designed to systematically benchmark multiple regressors and ensemble strategies. First, unimodal models are trained on visual features to establish baselines. Then, various fusion strategies are explored, including early concatenation, late score-level ensembles, and hybrid stacking. To improve generalization and reduce redundancy, PCA-based feature selection and cross-validation were applied. Finally, stacked meta-models are trained to integrate predictions from different feature sets, enabling nonlinear combinations that adapt to modality-specific strengths.

2 RELATED WORK

Research on video memorability prediction has expanded significantly in recent years, driven by the availability of benchmark datasets and shared evaluation campaigns. The VIDEM dataset introduced large-scale annotations of video memorability, enabling systematic comparisons across modeling approaches [1]. Early studies demonstrated that memorability is not solely a function of low-level visual features but also depends on semantic richness, narrative cues, and affective content [1]. Deep visual encoders have become central to feature extraction in this domain. Convolutional neural networks such as VGG [3] and ResNet [4] provide hierarchical representations of spatial structure, while more recent Vision Transformers (ViT) capture long-range dependencies and global context [5]. Complementary approaches have also leveraged multimodal signals, including audio, text, and even neural responses. For example, EEG-based studies have shown that brain activity during viewing can provide predictive cues for subsequent recall [6]. Beyond unimodal modeling, fusion strategies have been investigated to combine complementary modalities. Previous work has contrasted early fusion (feature concatenation) with late fusion (score-level ensembles), often finding that hybrid approaches yield more robust predictions [1]. Ensemble learning

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online
bhuvanaj@ssn.edu.in (J.Bhuvana); ramanan@ssn.edu.in (R.Mahendran); siddharth2312010@ssn.edu.in (S.Siddharth Chandrasekar); pragatheesh2312067@ssn.edu.in (J. Pragatheesh)



CEUR Workshop Proceedings (CEUR-WS.org)

techniques such as stacked generalization have proven effective in integrating diverse predictors, allowing nonlinear meta-models to exploit modality-specific strengths while mitigating weaknesses [7].

This work builds on these directions by systematically comparing visual encoders (ResNet, VGG, ViT, R3D) and evaluating ensemble-based fusion strategies for predicting both movie and commercial memorability. By using this approach within this body of research, it highlights the trade-offs between unimodal baselines, multimodal integration, and stacked ensembles in the context of MediaEval's shared benchmark.

3 APPROACH

This approach to predicting movie and commercial memorability is structured around a modular pipeline that integrates feature extraction, preprocessing, model training, fusion strategies, and evaluation. Each challenge subtask required a tailored design, but all shared the principle of combining complementary modalities with ensemble learning to improve reliability and predictive power.

3.1 Movie Memorability (Visual Features)

For the first subtask, the approach focuses on visual features extracted from video frames using ViT and R3D encoders. These models capture complementary aspects of video content, ViT encodes global semantic context using self-attention, while R3D captures spatiotemporal dynamics through 3D convolutions. After preprocessing with normalization and SelectKBest feature selection, two base regressors were trained, a Support Vector Regressor (SVR) with RBF kernel and an XGBoost regressor. Their predictions were combined using a stacked ensemble, with a Gradient Boosting Regressor as the meta-model. This configuration allowed to integrate predictions from diverse learners while mitigating overfitting. Evaluation was performed with 5-fold cross-validation, using Spearman correlation as the primary metric, along with MSE and R^2 for reliability assessment.

3.2 EEG-Based Recall Detection

The second subtask addressed EEG-based recall prediction, where the goal was to classify whether a video segment was recalled based on neural signals. ERP and ERSP features extracted from EEG recordings were processed and aligned with video stimuli. ERP signals were summarized using statistical descriptors such as mean, standard deviation, max, min, capturing temporal dynamics of brain responses. ERSP features were aggregated across frequency bands and epochs to encode spectral patterns. The combined feature set was used to train a Random Forest classifier with 100 trees. This model provided a strong baseline for recall detection, demonstrating that neural signals contain predictive cues for memorability.

3.3 Video Memorability with Multimodal Fusion

For the third subtask, the pipeline was extended to incorporate visual, textual, and metadata features. Visual features were extracted from AlexNet and VGG embeddings, chosen for their complementary spatial representations. Metadata fields such as duration, view counts, and engagement rate were combined with Whisper ASR transcripts, which were embedded using a SentenceTransformer model (all-MiniLM-L6-v2) [5] . These multimodal features were concatenated and reduced via Principal Component Analysis (PCA) to ensure compact and decorrelated representations. Then a stacked regressor with Ridge, SVR, and Gradient Boosting as base models, and Gradient Boosting as the meta-model was trained. This setup allowed us to integrate heterogeneous modalities into a unified prediction framework.

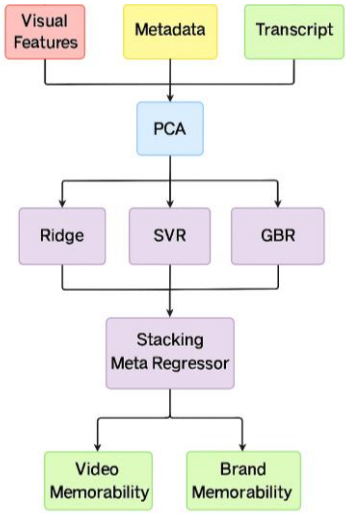

Figure 1: Overview of the multimodal memorability prediction pipeline.

Figure 1 depicts clear visual of the modeling pipeline showing how raw features flow through PCA, into base models (Ridge, SVR, Gradient Boosting), and finally into a stacked meta-model that outputs predictions for video and brand memorability.

3.4 Brand Memorability Prediction

The fourth subtask followed the same multimodal pipeline as Video memorability but targeted brand memorability instead of general video memorability. A separate stacked regressor trained on the same fused feature space was able to capture both content-level and brand-related signals.

3.5 Fusion and Evaluation Strategy

Across all subtasks, early fusion (feature concatenation), late fusion (score-level averaging), and hybrid stacking (meta-model integration) were systematically compared. Stacking consistently provided the most reliable improvements, as it allowed nonlinear integration of modality-specific strengths. For example, in Challenge 2.1, stacking improved Spearman correlation from -0.15 (Run 1, unimodal baseline) to 0.28 (Run 2, multimodal fusion with stacking), and reduced MSE from 0.031 to 0.021. Similarly, in Challenge 1.1, our stacked regressor achieved a Spearman correlation of 0.14, outperforming individual base models. All models were evaluated using Spearman rank correlation as the primary performance metric, with additional reporting of MSE and R^2. Cross-validation ensured robustness and reproducibility, and PCA (retaining 50 components) helped control feature dimensionality and reduce overfitting. This unified pipeline enabled us to tackle diverse memorability tasks with a consistent methodology, while adapting to the unique characteristics of each challenge.

4 RESULTS AND ANALYSIS

Models for all four subtasks in the MediaEval 2025 Predicting Movie and Video Memorability challenge were submitted. Below we summarize the performance of each run and analyze the outcomes in relation to the modeling strategies.

4.1 Movie Memorability Prediction

Table 1: **Result score for the Movie memorability prediction run**

Run ID	Spearman	Pearson	MSE
me25mem_SSN-CSE-CODECATALYSTS_challenge11_run1	0.14	0.163	0.069

Table 1 shows the first run for movie memorability used stacked regressors trained on ViT and R3D features. While the Spearman correlation (0.14) indicates moderate ranking agreement, the relatively high MSE suggests that the model struggled with precise score prediction. This may be due to limited modality coverage or insufficient temporal modeling. Future iterations could benefit from incorporating audio or text features and exploring transformer-based temporal encoders.

4.2 EEG-Based Recall Detection

Table 2: **Result score for the EEG based recall detection run**

Run ID	AUC
me25mem_SSN-CSE-CODECATALYSTS_challenge12_run1	0.47

Table 2 shows that the EEG-based recall model achieved an AUC of 0.47, slightly below the random baseline. This result highlights the challenge of modeling noisy EEG data, especially with limited training samples. Although ERP and ERSP features were created and used to train a Random Forest classifier, further improvements may require more sophisticated temporal alignment, artifact removal, or deep learning approaches tailored to EEG signals.

4.3 Multimodal Memorability Prediction

Table 3: **Result score for the challenge 2.1 run**

Run ID	Spearman	Pearson	MSE
me25mem_SSN-CSE-CODECATALYSTS_challenge21_run1	-0.15	-0.088	0.031
me25mem_SSN-CSE-CODECATALYSTS_challenge21_run2	0.28	0.32	0.021

Two runs were submitted for this subtask and the result score can be seen in Table 3. The first run yielded negative correlations (Spearman: -0.15, Pearson: -0.088), indicating poor generalization and possible feature misalignment. This run relied on unimodal features without proper fusion or dimensionality reduction, which likely led to overfitting and weak signal capture. In contrast, the second run using fused AlexNet and VGG features, metadata, and Whisper transcripts embedded via SentenceTransformer achieved a Spearman correlation of 0.28, MSE of 0.021, and a Pearson correlation of 0.32. Prior to modeling, we applied Principal Component Analysis (PCA) and retained 50 components based on cumulative explained variance. This validated the effectiveness of multimodal fusion and PCA-based dimensionality reduction. The improvement between runs reflects the importance of modality selection, preprocessing, and ensemble learning in predicting memorability.

4.4 Brand Memorability Prediction

Table 4: **Result score for the Brand memorability run**

Run ID	Spearman	Pearson	MSE
me25mem_SSN-CSE-CODECATALYSTS_challenge22_run1	-0.24	-0.204	0.024
me25mem_SSN-CSE-CODECATALYSTS_challenge22_run2	0.02	0.004	0.028

Brand memorability proved more difficult to model. From Table 4 the first run showed negative correlations, while the second run barely exceeded random performance. Despite using the same multimodal pipeline as in Multimodal Memorability, the weaker results suggest that brand memorability may depend on subtler cues—such as logo visibility, brand sentiment, or prior exposure that were not captured in our current feature set. Incorporating external brand knowledge graphs or fine-tuned sentiment models may improve future performance.

4.5 Summary and Insights

Stacked ensembles improved reliability in multimodal prediction but were less effective in movie memorability and the EEG-based recall model, likely due to limited modality diversity or weak signal quality. For multimodal memorability, the fusion of visual, metadata, and transcript features clearly outperformed unimodal baselines. EEG modeling, however, remains challenging due to noise and data sparsity; future work should explore temporal deep models and domain-specific preprocessing. Similarly, brand memorability may require richer semantic and contextual features beyond what standard encoders can provide.

5 CONCLUSIONS

In this paper, we present a unified pipeline for predicting video and commercial memorability across four MediaEval 2025 subtasks. Our approach combines deep visual encoders (such as ViT, R3D, AlexNet, and VGG), metadata, and transcript embeddings with ensemble learning techniques, including stacked regressors and Random Forest classifiers. For EEG-based recall detection, we engineered ERP and ERSP features to capture the neural correlation of memory.

Across the subtasks, our work demonstrates that multimodal fusion and stacking strategies can improve predictive reliability, particularly when modalities are complementary. While these models achieved promising results in multimodal memorability, other subtasks—such as EEG-based recall and brand memorability—highlighted limitations in signal quality, feature coverage, and generalization. Future work will explore transformer-based temporal modelling, uncertainty-aware ensembles, and domain-specific feature engineering for EEG and brand semantics. Finally, to improve reproducibility, we are releasing our modular code and benchmarking scripts to support the broader research community.

ACKNOWLEDGMENTS

We would like to thank the organizers of the MediaEval 2025 Predicting Movie and Video Memorability task for providing the datasets, evaluation framework, and collaborative environment. We acknowledge the support of Dr. J. Bhuvana whose research guidance and the Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering whose computational resources enabled the completion of this work.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

REFERENCES

[1] Iván Martín-Fernández, Mihai Gabriel Constantin, Claire-Hélène Demarty, Manuel Gil-Martín, Sebastian Halder, Bogdan Ionescu, Rukiye Savran Kiziltepe, Ana Matran-Fernandez, and Alba G. Seco de Herrera. 2025. Overview of the MediaEval 2025 Predicting Movie and Commercial Memorability Task. In Working Notes Proceedings of the MediaEval 2025 Workshop. CEUR Workshop Proceedings, Dublin, Ireland and Online.

[2] Mohammad Soleymani, et al. 2017. A Multi-Modal Approach to Video Memorability Prediction. In Proceedings of the ACM International Conference on Multimedia (MM '17). ACM, New York, NY, USA. DOI: https://doi.org/10.1145/3123266.3127909

[3] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In International Conference on Learning Representations (ICLR '15). arXiv:1409.1556

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16). IEEE, 770–778. DOI: https://doi.org/10.1109/CVPR.2016.90

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations (ICLR '21). arXiv:2010.11929

[6] Aditya Khosla, Constance Bainbridge, Antonio Torralba, and Aude Oliva. 2013. Predicting Video Memorability from EEG Signals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '13). IEEE, 29–34. DOI: https://doi.org/10.1109/CVPRW.2013.12

[7] David H. Wolpert. 1992. Stacked Generalization. Neural Networks 5, 2 (1992), 241–259. DOI: https://doi.org/10.1016/S0893-6080(05)80023-1

[8] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP '19). Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/D19-1410

[9] Daniel Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV '15). IEEE, 4489–4497. DOI: https://doi.org/10.1109/ICCV.2015.510

[10] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 785–794. DOI: https://doi.org/10.1145/2939672.2939785

[11] Ana Matran-Fernandez and Sebastian Halder. 2024. An EEG Dataset to Study Neural Correlates of Audiovisual Long-Term Memory Retrieval. Research Square Preprint. DOI: https://doi.org/10.21203/rs.3.rs-7066609/v1

[12] Ana Matran-Fernandez and Sebastian Halder. 2024. Essex EEG Movie Memory Dataset. OpenNeuro. DOI: https://doi.org/10.18112/openneuro.ds006142.v1.0.1

[13] Rukiye Savran Kiziltepe, S. Sahab, R. V. Santana, F. Doctor, K. Paterson, D. Hunstone, and Alba G. Seco de Herrera. 2025. VIDEM: VIDeo Effectiveness and Memorability Dataset. In Proceedings of the 18th International Work-Conference on Artificial Neural Networks (IWANN 2025), A Coruña, Spain.

[14] Lorin Sweeney, Graham Healy, and Alan F. Smeaton. 2023. Diffusing Surrogate Dreams of Video Scenes to Predict Video Memorability. In Working Notes Proceedings of the MediaEval 2022 Workshop, CEUR Workshop Proceedings, Bergen, Norway. URL: https://ceur-ws.org/Vol-3583/paper52.pdf

[15] Hitesh Si, Shubham Singh, Yash K. Singla, Anirban Bhattacharyya, Vishnu Baths, Chao Chen, Rajiv R. Shah, and Bhargav Krishnamurthy. 2025. Long-Term Ad Memorability: Understanding & Generating Memorable Ads. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV '25). IEEE, 5707–5718. DOI: https://doi.org/10.1109/WACV61041.2025.00557