Multimodal Feature Fusion for Video and Brand Memorability

Elif Nur Tekay[1,*,†], Irem Azra Isleyen[1,†], Murat Karakus[1,†] and Rukiye Savran Kiziltepe[1,†]

[1]*Department of Software Engineering, Ankara University, Ankara, 06830, Türkiye*

Abstract

This paper presents ANLAM-NET Lab's participation in the *MediaEval 2025 Predicting Movie and Commercial Memorability* task. We developed a multi-stream regression architecture that combines multimodal deep visual and textual representations. The visual stream integrates Convolutional Neural Network (CNN) and Transformer architectures—including AlexNet, DenseNet121, EfficientNetB3, ResNet50, VGG, ViT, and R3D—together with CLIP-text embeddings derived from video transcripts, titles, and descriptions, as well as CLIP-image features extracted from thumbnails and multi-frame inputs. To jointly predict video and brand memorability, we explored Gated Multi-layer Perceptron (MLP) fusion, Transformer-based token fusion, and Random Forest (RF) architectures. Experimental results show that the RF model achieved the highest test performance, while multimodal fusion strategies—particularly gated fusion and Transformer architectures—consistently outperformed unimodal and hand-crafted feature-based models.

1. Introduction

Predicting media memorability is a challenging problem in multimedia research with broad implications for advertising, human–computer interaction, and cognition. The MediaEval 2025 *Predicting Movie and Commercial Memorability* task [1] addresses this challenge through a benchmark for the joint prediction of video and brand memorability. Subtask 2 focuses on long-term memorability of commercial videos using the *VIDEM* dataset[2], which contains 424 annotated commercial clips with video and brand memorability scores. Systems are required to jointly estimate both scores, encouraging multimodal learning that combines visual, textual, and semantic cues relevant to brand recall.

Early work showed that image memorability is consistent across observers and computationally predictable [3, 4]. Subsequent studies extended this to video, emphasizing semantic and affective cues [5] and inspiring multimodal approaches benchmarked through MediaEval tasks [6, 7]. Recent vision–language models, notably CLIP [8], advanced cross-modal prediction; Nguyen *et al.* [9] demonstrated the effectiveness of CLIP-based representations and three-frame sampling. Beyond early and late fusion, gated fusion adaptively weights modalities and tolerates missing inputs [10].

Building on these findings, we explore multimodal regression architectures integrating CLIP-based textual, thumbnail, and multi-frame features with deep visual representations. We systematically compare fusion strategies (early, late, gated) across regression models including Ridge, Multi-layer Perceptron (MLP), Random Forest (RF), XGBoost, and a Transformer-based

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

*Corresponding author.

†These authors contributed equally.

23291762@ogrenci.ankara.edu.tr (E. N. Tekay); 22291007@ogrenci.ankara.edu.tr (I. A. Isleyen); mrtkarakus@ankara.edu.tr (M. Karakus); rukiyekiziltepe@ankara.edu.tr (R. Savran Kiziltepe)

Table 1
Model architectures and key configuration details.

Models	Regression Models	Features	Input Modalities	Aggregation Techniques	Fusion Techniques
Model-1	MLP	AlexNet, ResNet50, VGG, ViT, R3D	Visual (5 CNN); CLIP-Text (STT+title+desc); CLIP-Thumbnail; CLIP-Image (first/middle/last)	Mean+Std per modality	Gated
Model-2	MLP	AlexNet, ResNet50, VGG, ViT, R3D	Visual (5 CNN); CLIP-Text; CLIP-Thumbnail	PCA-256	Gated
Model-3	MLP	None	Visual (single 512-D); CLIP-Text; CLIP-Thumbnail	Global PCA or Mean+Std $\rightarrow$ PCA	Gated
Model-4	RF	AlexNet, EfficientNetB3, ResNet50, VGG, ViT, R3D	Visual (selected CNN architectures)	Concat / Mean+Std / PCA-256	Early (feature-level concat)
Model-5	Transformer Fusion	AlexNet, DenseNet121, EfficientNetB3, ResNet50, VGG, ViT, R3D	Tokens: $7\times$CNN, $1\times$CLIP-Text, $1\times$CLIP-Thumbnail, $1\times$CLIP-Image ($\approx$9 tokens)	Per-modality token	Transformer encoder

fusion model. Our results indicate that gated fusion combined with mean+std aggregation provides a robust framework for predicting both video and brand memorability.

2. Methodology

This section details the multimodal methodology developed for the MediaEval 2025 *Predicting Movie and Commercial Memorability* task. We describe the preprocessing steps applied to the *VIDEM* dataset, followed by the feature extraction, fusion, and modeling strategies employed for predicting both video and brand memorability.

2.1. Feature Extraction and Fusion Strategies

Our multimodal framework integrates three main modalities: visual, textual, and image-based features.

Visual features: Deep visual features provided by the organizers—comprising multiple CNN and Transformer architectures (AlexNet, DenseNet121, EfficientNetB3, ResNet50, VGG, ViT, R3D) alongside handcrafted descriptors (HSV, RGB and LBP)—were processed under three aggregation schemes (concatenation, PCA-256, and mean+std pooling), while thumbnails and first/middle/last video frames were encoded using CLIP-image (512-D and zero if missing) and mean+std pooled into a 1024-D representation to capture temporal context.

Textual features: Transcripts, titles, and descriptions were concatenated and encoded with CLIP-text into 512-D embeddings.

We compared three fusion strategies: early fusion (feature concatenation), late fusion (decision-level combination), and gated fusion [11]. In gated fusion, each modality is projected to a shared hidden space ($h = 768$) and reweighted by a learnable non-negative gate before normalization: allowing adaptive weighting of modalities and robustness to missing or noisy data.

2.2. Model Architectures

Table 1 lists the implemented multimodal regression models jointly predicting video and brand memorability, all using (i) visual features, (ii) 512-D CLIP-text embeddings from transcripts, titles, descriptions; and (iii) 512-D CLIP-image embeddings from thumbnails. Figure 1 illustrates the multimodal data flow and methodology structure followed in this study.

The first three models are based on a MLP with a gated fusion backbone. Model-1 combines mean and standard deviation pooled CNN features with CLIP embeddings from text, thumbnails, and the first, middle, and last video frames. Model-2 is similar but uses PCA-compressed visual

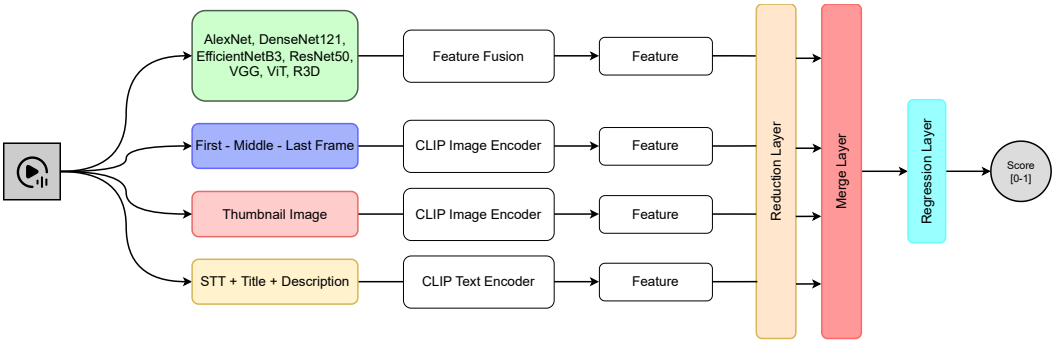

Figure 1: Multimodal architecture combining CNN, CLIP-image, and CLIP-text features for video and brand memorability prediction.

features instead of the 3-frame CLIP embeddings. Model-3 further simplifies the visual input by merging all visual features into a single 512-D vector before fusion. In contrast, Model-4 employs a Transformer encoder to model long-term dependencies which represent each modality—including provided CNN features, CLIP-text, CLIP-thumbnail, and CLIP-image—as individual tokens processed through the encoder. Finally, Model-5 is used as a non-neural baseline, using a RF ensemble to combine selected visual features through early, feature-level concatenation.

Training followed two regimes: *validation mode* (early stopping, patience=10–20, key=mean Spearman) and *final-fit mode* (train+val merged for 20–45 epochs). The RF baseline used grid search and early stopping on validation MSE. All models saved the best checkpoints and produced standardized CSV outputs.

3. Results and Discussion

This section summarizes the quantitative results and main experimental findings regarding the effect of fusion strategies, feature representations, and model types on memorability prediction.

As the *VIDEM* test labels were withheld, a stratified 90/10 validation split of the training set was used for model development and hyperparameter tuning, allowing objective performance comparison. The validation and test results are summarized in Table 2.

Table 2
Validation and test results for both tasks: Challenge 2.1 (Video) and Challenge 2.2 (Brand) Memorability.

		Challenge 2.1 (Video Memorability)			Challenge 2.2 (Brand Memorability)		
	Model	**SRCC**	**PCC**	**MSE**	**SRCC**	**PCC**	**MSE**
Validation	**MLP + Gated Fusion + 3-Frame CLIP**	**0.401**	**0.393**	**0.023**	0.365	0.358	0.019
	MLP + Gated Fusion, PCA-256 visual	0.278	0.344	0.367	**0.376**	**0.351**	**0.266**
	MLP + Single-Vector Visual	0.284	0.256	0.021	0.227	0.270	0.022
	RF	0.345	0.338	0.012	0.099	0.104	0.021
	Transformer Fusion	0.198	0.217	0.018	-0.020	0.014	0.031
Test	MLP + Gated Fusion + 3-Frame CLIP	0.057	0.071	0.039	-0.011	0.049	0.035
	MLP + Gated Fusion, PCA-256 visual	-0.216	-0.200	0.049	-0.089	-0.101	0.042
	MLP + Single-Vector Visual	-0.071	-0.134	0.027	0.086	0.067	0.026
	RF	**0.203**	**0.254**	**0.026**	**0.124**	**0.145**	**0.025**
	Transformer Fusion	-0.038	0.044	0.028	0.017	0.026	0.029

On validation, Model-1 (MLP + Gated Fusion + 3-Frame CLIP) achieved the best video memorability results (Spearman 0.401, Pearson 0.393), confirming the benefit of gated fusion and temporal context. For brand memorability, the superior performance of Model-2 (MLP +

Gated Fusion, PCA-256 visual) suggests that more compact visual representations may facilitate learning in brand-specific scenarios. RF showed consistent yet slightly lower validation results.

On the official test set, a noticeable performance drop was observed in fusion models, which may be attributed to their sensitivity to potential distribution shifts between the training and test datasets. In contrast, the RF model yielded relatively higher correlations and lower MSE values across both tasks.

Across experiments, mean+std pooling appeared to be more effective than PCA-256 and direct concatenation. The inclusion of CLIP-text and CLIP- thumbnails features provided consistent improvements over visual-only models, while gated fusion was observed to outperform early fusion, particularly in cases with missing modalities.

We attribute the gap between validation and official test performance to several data- and modality-related factors: (i) limited sample size (424 videos) relative to the dimensionality of the feature space, which increases overfitting risk for neural models; (ii) missing modalities, notably inaccessible thumbnails, reducing the benefit of CLIP-image and weakening gated fusion's ability to calibrate contributions; (iii) distribution shift between train/validation and test (content style, brand frequency, and editing patterns); and (iv) potential label variability in brand memorability due to subjective brand exposure and recall. These factors collectively depress generalization even when validation metrics are competitive.

4. Conclusion

This study presented the multimodal regression architecture for long-term video and brand memorability and systematically evaluated different regression models, fusion strategies, and modality combinations. While MLP and Transformer models with gated fusion and multi-frame CLIP features performed best on validation (e.g., Model-1, Spearman 0.401), they underperformed on the official test set, indicating overfitting and sensitivity to distribution shift. The more stable test results of the RF model suggest that simpler, variance-reducing methods may offer advantages in terms of generalization under limited or incomplete data conditions. Future work will model temporal dynamics more richly using Transformer-based video models (e.g., TimeSformer, VideoMAE) and incorporate attention to detect the brand appears. We also plan to fine-tune Henry model on *VIDEM* to test task-specific and cross-domain generalization.

Declaration on Generative AI

During the preparation of this work, the authors used Gen-AI to improve the writing, check grammar and overall linguistic clarity. All technical content, experimental design, analysis, and interpretations were fully conceived, implemented, and critically reviewed by the authors. The authors take full responsibility for the originality and accuracy of the final manuscript.

Acknowledgements

This work was supported by TÜBİTAK under the 2209-A (Project No: 1919B012468481). This work received support from the 2224-B - Grant Program for Participation in Scientific Meetings within the Country.

References

[1] I. Martín-Fernández, 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 media-eval 2025 predicting movie and commercial memorability task, in: Proc. of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.

[2] R. S. Kiziltepe, S. Sahab, R. V. Santana, F. Doctor, K. Paterson, D. Hunstone, A. G. S. de Herrera, VIDEM: VIDeo effectiveness and memorability dataset, in: I. Rojas, G. Joya, A. Catala (Eds.), Advances in Computational Intelligence, Springer Nature Switzerland, Cham, 2025, pp. 41–54.

[3] 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.

[4] A. Khosla, A. S. Raju, A. Torralba, A. Oliva, Understanding and predicting image memorability at a large scale, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2390–2398.

[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/CVF International Conference on Computer Vision (ICCV), 2019.

[6] M. M. A. Usmani, S. Zahid, M. A. Tahir, Modelling of Video Memorability using Ensemble Learning and Transformers, in: Proc. MediaEval 2022 Multimedia Benchmark Workshop (MediaEval'22) – Working Notes, Bergen, Norway and Online, 2023.

[7] S. Azzakhinini, O. B. Ahmed, C. Fernandez-Maloigne, Video Memorability Prediction using Deep Features and Loss-based Memorability Distribution Estimation, in: Proc. MediaEval 2022 Multimedia Benchmark Workshop (MediaEval'22) – Working Notes, Bergen, Norway, 2022.

[8] 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: International Conference on Machine Learning, 2021.

[9] M.-Q. Nguyen, M.-H. Trinh, H.-G. Bui, K.-T. Vo, M.-T. Tran, T.-P. Tran, H.-D. Nguyen, Selab-hcmus at mediaeval 2023: A cross-domain and subject-centric approach towards the memorability prediction task, in: Proceedings of the MediaEval 2023 Workshop, 2023.

[10] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, X. Peng, Smil: Multimodal learning with severely missing modality, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 2302–2310.

[11] J. Arevalo, T. Solorio, M. Montes-y Gómez, F. A. González, Gated multimodal networks, Neural Computing and Application 32 (2020) 10209–10228.