A CLIP-Based Visual-Semantic Embedding Approach for Commercial Memorability Prediction

Elif Nur Tekay[1,*,†], Irem Azra Isleyen[2,†], Rukiye Savran Kiziltepe[2,†] and Murat Karakus[2,†]

[1]Department of Software Engineering, Malatya Turgut Ozal University, Malatya, Türkiye
[2]Department of Software Engineering, Ankara University, Ankara, 06830, Türkiye

Abstract

This paper presents our participation in Subtask 2 (Commercial/Ad Memorability) of the *MediaEval 2026: Predicting Movie and Commercial Memorability* task, focusing on the prediction of video and brand memorability using the *VIDEM* commercial video dataset. We propose a visual-semantic regression framework that combines CLIP-based frame embeddings, thumbnail embeddings, semantic representations derived from textual descriptions of uniformly sampled video frames, and metadata features. The textual descriptions are encoded using the CLIP text encoder, enabling semantic representations that are aligned with visual content. To investigate the contribution of different features, we evaluate CatBoost, Random Forest (RF), and transformer-based visual-semantic models under several feature combinations. On the official test set, CatBoost combined with interaction-based Principal Component Analysis (interaction-PCA) over thumbnail, textual-description, and metadata features achieved the best performance for video memorability prediction, obtaining a Spearman's Rank Correlation Coefficient (SRCC) of 0.279 and a Pearson Correlation Coefficient (PCC) of 0.258. For brand memorability prediction, the transformer-based visual-semantic model using textual-description and frame-image embeddings achieved the strongest ranking performance, reaching an SRCC of 0.254 and a PCC of 0.266.

1. Introduction

Predicting how well media content is retained in viewers' memory is an important challange in multimedia analysis and advertising research. Because memorability is influenced by both perceptual characteristic and high-level semantic information, effective prediction requires representations that capture visual appearance and semantic content jointly. In commercial videos, this challenge is further complicated by the fact that a memorable advertisement does not necessarily result in strong recall of the featured brand. The *MediaEval 2026 Predicting Movie and Commercial Memorability* task addresses this problem providing the *VIDEM* dataset, which consists of 424 commercial videos annotated with both video and brand memorability scores [1, 2]. In this work, we participated in Subtask 2, where the goal is to predict two targets: video memorability and brand memorability.

Previous studies on image and video memorability have shown that certain visual content is remembered consistently across viewers and that memorability can be predicted using computational models [3, 4]. Large-scale datasets such as VideoMem [4] and Memento10k [5] further highlighted the importance of semantic content and temporal dynamics in memorability prediction. In addition, topic-oriented textual representations have been shown to achieve

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
*Corresponding author.
†These authors contributed equally.
02220201053@ozal.edu.tr (E. N. Tekay); 22291007@ogrenci.ankara.edu.tr (I. A. Isleyen); rukiyekiziltepe@ankara.edu.tr (R. Savran Kiziltepe); mrtkarakus@ankara.edu.tr (M. Karakus)

performance comparable to that of visual deep-learning models for video memorability prediction [6]. Research conducted within the MediaEval benchmark has explored a wide range of approaches, including visual, textual, temporal, and metadata-based features, as well as feature fusion strategies, ensemble methods, traditional machine learning regressors, neural networks, and hybrid architectures [7]. Collectively, these studies suggest that video memorability cannot be fully explained by a single modality and that combining visual and semantic information remains a promising direction for improving prediction performance.

More recently, textual descriptions derived from visual content have been increasingly used to capture richer semantic information for memorability prediction. CLIP-based representations, in particular, have become a popular choice for aligning visual and textual information within a shared embedding space [8]. In the context of advertisement memorability, the MindMem framework demonstrated that representations extracted from large-scale foundation models can effectively support memorability prediction [9]. Similarly, recent studies have investigated semantic representations generated from visual content for video memorability regression. For example, Qwen-VL was fine-tuned using a parameter-efficient adaptation strategy for memorability prediction on the Memento10k dataset [10].

In our previous workshop paper, the provided visual features were combined with CLIP-based textual, thumbnail, and frame-level visual representations using MLP-based gated integration, Random Forest (RF), XGBoost, and Transformer-based regression approaches [11]. In this work, we retain the CLIP-based textual and thumbnail representations while extending the frame-level visual representation from three key frames (first, middle, and last) to 8 uniformly sampled frames. The feature space is enriched with semantic embeddings derived from textual descriptions generated for the sampled frames, together with metadata features. Using these representations, we evaluate CatBoost, RF, and Transformer Fusion models under different feature fusion settings, including interaction-based Principal Component Analysis (interaction-PCA), PCA/mean-std representations, and feature concatenation. Video memorability and brand memorability are analyzed as separate prediction targets.

2. Methodology

This section presents the visual-semantic framework developed for the Subtask 2 (Commercial/Ad Memorability) of the *Mediaeval2026 Predicting Movie and Commercial Memorability* [1]. The overall architecture of the proposed framework is shown in Figure 1.

2.1. Visual and Semantic Feature Extraction

For each video, we sampled 8 frames at uniformly distributed temporal positions to obtain a compact representation of the visual content. These frames were used to extract both semantic and visual features. For semantic representation, the sampled frames were processed with *Qwen2.5* to generate textual descriptions of the visible scene, objects, actions, and contextual elements. The generated descriptions were then encoded using the CLIP text encoder, producing 512-dimensional embeddings that represent high-level visual semantics.

In parallel, the same sampled frames were encoded directly using the CLIP image encoder to obtain frame-level visual embeddings. Thumbnail images were also encoded with the CLIP image encoder and treated as a separate feature source. In addition, metadata features associated with the videos were included.

Overall, four feature groups were used: textual-description embeddings, CLIP-based frame-image embeddings, CLIP-based thumbnail embeddings, and metadata features. These feature

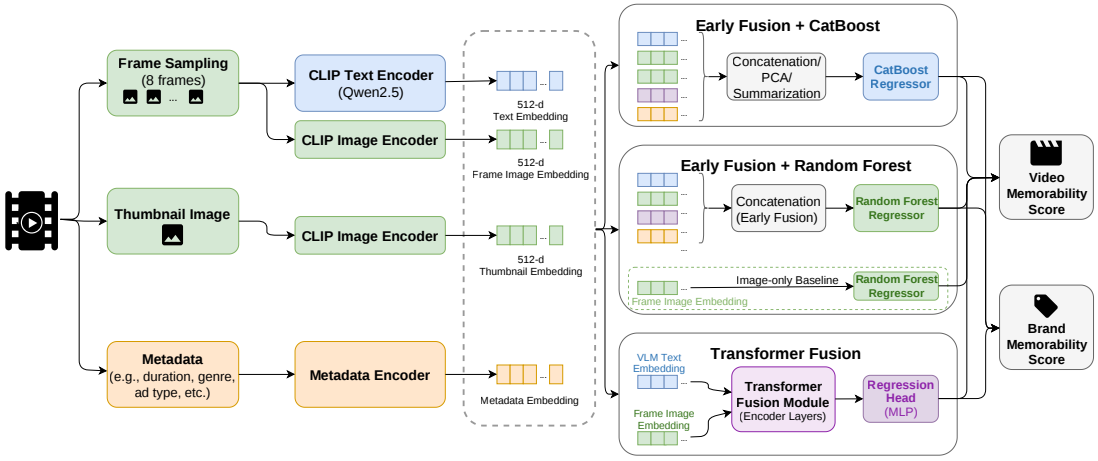

Figure 1: Visual-semantic feature fusion architecture for video and brand memorability prediction.

groups were used in different combinations to analyze their individual and complementary contributions to video and brand memorability prediction.

2.2. Feature Fusion Strategies and Regression Models

We evaluated three modelling approaches: CatBoost, RF, and Transformer Fusion. In the CatBoost and RF experiments, selected feature vectors were concatenated into a single feature representation before regression. For CatBoost, different combinations of textual-description embeddings, frame-image embeddings, thumbnail embeddings, and metadata features were evaluated to examine the contribution of each feature group.

For the RF experiments, two configurations were used. The first configuration served as an image-only baseline, while the second combined textual-description embeddings, frame-image embeddings, and thumbnail embeddings. This comparison was designed to evaluate whether semantic descriptions and thumbnail-based features provide additional predictive information beyond frame-level visual features alone.

We also implemented a transformer-based visual-semantic model to process feature-specific representations more explicitly. Instead of directly concatenating all features, textual-description embeddings and frame-image embeddings were used as separate inputs. The transformer encoder was then used to model interactions between semantic and visual representations before predicting memorability scores. Model selection was based on validation performance, and the selected configurations were used to generate the official test predictions.

3. Results and Discussion

Table 1 summarizes the validation and official test results for Challenge 2.1 and Challenge 2.2. Performance was evaluated using Spearman's Rank Correlation Coefficient (SRCC), Pearson Correlation Coefficient (PCC), and Mean Squared Error (MSE). Since memorability prediction requires estimating continues target scores while preserving the relative ranking of videos, SRCC is used as one of the primary evaluation metrics for this task.

For Challenge 2.1, the best video memorability performance was obtained with CatBoost using thumbnail, textual-description, and metadata features represented with interaction-PCA. It reached 0.279 SRCC, 0.258 PCC, and 0.025 MSE on the official test set. This result suggests

Table 1
Validation and official test results for video (Challenge 2.1) and brand memorability (Challenge 2.2).

Task	Set	Model / Feature Setting	SRCC	PCC	MSE
Challenge 2.1	Validation	**CatBoost + interaction PCA (Thumb + Text-desc. + Metadata)**	**0.351**	**0.446**	**0.017**
		RF + Text-desc./Image/Thumbnail	0.197	0.140	0.021
		RF image-only baseline	0.112	0.126	0.022
		Transformer Fusion (Text-desc. + Image)	0.171	0.128	0.021
	Test	**CatBoost + interaction PCA (Thumb + Text-desc. + Metadata)**	**0.279**	**0.258**	**0.025**
		RF + Text-desc./Image/Thumbnail	0.139	0.187	0.026
		RF image-only baseline	0.020	0.117	0.026
		Transformer Fusion (Text-desc. + Image)	0.148	0.148	0.027
Challenge 2.2	Validation	RF + Text-desc./Image/Thumbnail	0.264	0.159	0.025
		Transformer Fusion (Text-desc. + Image)	**0.286**	**0.271**	**0.023**
	Test	RF + Text-desc./Image/Thumbnail	0.030	0.040	0.026
		Transformer Fusion (Text-desc. + Image)	**0.254**	**0.266**	**0.025**

that semantic descriptions, thumbnail-based visual summaries, and metadata provide useful cues for video memorability. In contrast, the image-only RF baseline remained at 0.020 SRCC on the official test set, suggesting that frame-level visual embeddings alone provide limited predictive capacity for this task.

For Challenge 2.2, the transformer-based visual-semantic model achieved the best brand memorability performance among the reported visual-semantic configurations, with 0.254 SRCC, 0.266 PCC, and 0.025 MSE on the official test set. This result suggests that explicitly modelling interactions between semantic descriptions and frame-level visual representations is useful for brand memorability prediction. The RF model using text, image, and thumbnail features achieved lower official test SRCC, indicating that direct feature-level concatenation may have limited generalization capability for brand memorability.

4. Conclusion

This study presented a visual-semantic framework for predicting video and brand memorability on the *VIDEM* dataset realased in Subtask 2 (Commercial/Ad Memorability) of the MediaEval 2026: Predicting Movie and Commercial Memorability task. The results suggest that video and brand memorability benefit from different modelling strategies, with CatBoost performing best for video memorability and the transformer-based model achieving the strongest results for brand memorability. Future work will explore improved temporal modelling, brand-aware attention mechanisms, and feature ablation analyses.

Declaration on Generative AI

During the preparation of this work, Gen-AI was used only for language refinement and grammar checking. All technical content, experiments, analysis, and interpretations were conceived, implemented, and reviewed by the authors, who take full responsibility for the final manuscript.

Acknowledgements

This work was supported by TÜBİTAK under the 2209-A (Project No: 1919B012468481).

References

[1] I. Martín-Fernández, M. G. Constantin, C.-H. Demarty, M. Gil-Martín, S. Halder, B. Ionescu, R. Savran Kiziltepe, A. Matran-Fernandez, A. G. 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] R. Savran Kiziltepe, S. Sahab, R. Valladares Santana, F. Doctor, K. Paterson, D. Hunstone, A. G. Seco de Herrera, VIDEM: VIDeo effectiveness and memorability dataset, in: I. Rojas, G. Joya, A. Catala (Eds.), Advances in Computational Intelligence, volume 16008 of *Lecture Notes in Computer Science*, Springer Nature Switzerland, Cham, 2026, 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, 2011, pp. 145–152.

[4] 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, 2019, pp. 2531–2540.

[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: Computer Vision – ECCV 2020, volume 12361 of *Lecture Notes in Computer Science*, Springer, Cham, 2020, pp. 223–240.

[6] R. Kleinlein, C. Luna-Jim'enez, D. Arias-Cuadrado, J. Ferreiros, F. Fern'andez-Mart'inez, Topic-oriented text features can match visual deep models of video memorability, Applied Sciences 11 (2021) 7406.

[7] 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. G. Seco de Herrera, A review of computational memorability: A benchmark framework, International Journal of Computer Vision 134 (2026) 298.

[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: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763.

[9] S. Asgarian, Q. Jetha, J. Jeon, MindMem: Multimodal for predicting advertisement memorability using LLMs and deep learning, arXiv preprint arXiv:2502.18371 (2025).

[10] 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.

[11] E. N. Tekay, I. A. Isleyen, M. Karakus, R. Savran Kiziltepe, Multimodal feature fusion for video and brand memorability, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.