Visual Transformers and Contextual Features for Movie Memorability Estimation

Aashutosh Ganesh[1,*], Mirela Popa[1] and Nava Tintarev[1]

[1]*Maastricht University, The Netherlands*

Abstract

This paper presents our approach for Challenge 1.1, Long-Term Movie-Clip Memorability Estimation at MediaEval 2026. Our method builds upon the best approaches from the previous edition, while integrating the “neutral/typical" meta-data labels provided in this year’s task. We propose five different late fusion strategies, that combine our proposed transformer architecture with a variety of frame-level and video-level features, as well as meta-data labels. Two of our approaches, which separately fuse frame-level feature transformers with video-level predictors and neutral-typical labels, achieved the strongest performance on the test set, obtaining Spearman correlation coefficients of 0.34 and 0.42 respectively. These results demonstrate the importance of contextual information provided by the “neutral/typical" labels for movie memorability prediction. However, understanding the specific cues that determine whether a clip is classified as neutral or typical is an open challenge and warrants further investigation.

1. Introduction

A longstanding question in film-making has been; “What factors contribute towards the memorability of a movie clip?" While a film’s visual style can make certain shots more striking than others, memorability may extend beyond purely visual characteristics. Contextual elements such as famous actors or plot relevant scenes could make a particular scene more recognizable. To answer this question, Challenge 1.1 in this year’s edition of MediaEval’s memorability track [1], provides a dataset of movie clips accompanied by memorability annotations and additional labels indicating whether a clip is “typical" or “neutral". These labels indicate whether the source film can be recognized from contextual cues present in the clip, offering an opportunity to study the role of contextual information in memorability prediction.

For **Challenge 1.1**, our approach builds upon the two best approaches from the previous edition, which utilized a late fusion strategy [2] and a frame-wise transformer model [3]. We extend the latter by using frame-level visual features as input to a video transformer and introducing a CLS-token pooling strategy to aggregate temporal information for memorability prediction. Our late-fusion framework consists of four components: two instances of the proposed transformer architecture, each operating on a different set of frame-level visual features; a multilayer perceptron (MLP) that processes video-level motion features; and a predictor based on the neutral/typical labels provided for each video. We evaluate several fusion configurations by combining the predictions of these components through a weighted sum. Our results show that fusing transformer-based frame-level representations with either motion features or neutral/typical contextual labels improves upon the best-performing approach from

MediaEval’26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online

*Corresponding author.

Aashutosh.Ganesh@maastrichtuniversity.nl (A. Ganesh); mirela.popa@maastrichtuniversity.nl (M. Popa); n.tintarev@maastrichtuniversity.nl (N. Tintarev)

0000-0002-6449-1158 (M. Popa); 0000-0003-1663-1627 (N. Tintarev)

CEUR Workshop Proceedings (CEUR-WS.org)

the previous edition. Specifically, the previous state-of-the-art method achieved a Spearman correlation of 0.332 on the test set [2], whereas our best-performing configurations achieved correlations of 0.344 and 0.427, respectively. These findings highlight the value of contextual information for movie memorability prediction. However, understanding which characteristics lead a clip to be classified as neutral or typical remains an open research question.

2. Approach: Challenge 1.1

This section is organized as follows. Section 2.1 introduces the feature extractors used in our approach, while Section 2.2 describes the neural network architectures developed for the memorability prediction task.

2.1. Features

Visual features. Consider a clip $V = \{f_1, f_2, \dots f_N\}$, f_i is a video frame at index i and N is the number of sampled frames. A feature extractor F_V converts each frame to representation $\hat{f_i}$ producing $X = F_V(V) = \{\hat{f_1}, \hat{f_2}, \dots \hat{f_N}\}$. For feature extraction, we employ the vision–language model Siglip2 [4] (provided by Huggingface [5]) returning $\mathbb{R}^{1\times768}$ per frame. Our choice is motivated by prior work demonstrating the effectiveness of text-grounded visual representations for video memorability prediction [6], including the use of SigLip features [3]. Furthermore, SigLip2 improves upon the original SigLip model across a variety of downstream tasks, including image-text retrieval. To complement SigLip2's high-level semantic representations, we also extract self-supervised visual features using DinoV2 [7], which likewise produces a feature vector $\mathbb{R}^{1\times768}$ for each frame. DinoV2 has been shown to capture lower-level visual characteristics more effectively than vision-language models such as SigLip2 [8], making it a complementary source of information. We uniformly sample $N = 8$ frames per video, resulting in a feature matrix $\hat{F} \in \mathbb{R}^{8\times768}$ for both the SigLip2 and DinoV2 feature extractors.

Motion features. We utilize the TimeSformer [9] architecture, pre-trained on the Kinetics dataset to extract motion level features from the video. Consider a clip $V = \{f_1, f_2, \dots f_N\}$, where f_i is a video frame at index i and N is the number of sampled frames. A feature extractor F_M converts the sampled frames from the video to a representation $\hat{F_m}$ producing $\hat{F_m} = F_M(V)$. The resulting vector is $\hat{F_m} \in \mathbb{R}^{1\times512}$. This choice was motivated by the fact that newer motion feature extractors have not yet been evaluated on the MovieMem dataset, despite prior work demonstrating the effectiveness of motion-based features for movie memorability prediction [10].

Neutral and Typical Feature This edition of the challenge provides 'neutral' and 'typical' labels for each video, indicating whether the clip contains contextual cues that make the source film recognizable. We represent this information as a binary feature, $F_{nt} \in 0, 1$, where 0 corresponds to a neutral clip and 1 corresponds to a typical clip. We incorporate this feature into our late-fusion framework to investigate the contribution of contextual information to movie memorability prediction.

2.2. Model Architectures

Video Transformer A transformer architecture P (depicted in Fig 1) maps the sequential feature representations $\hat{F}$ to a memorability score $m_V = P(\hat{F}) \in [0, 1]$. Similar to previous work, the architecture consists of four components with several modifications: (1) sinusoidal positional encoding [11] are added to the input sequence, (2) 5 stacked transformer blocks, each containing multi-headed self-attention [11] (8 heads with 64 dimensions), $\hat{F}$ is projected to $H \in$

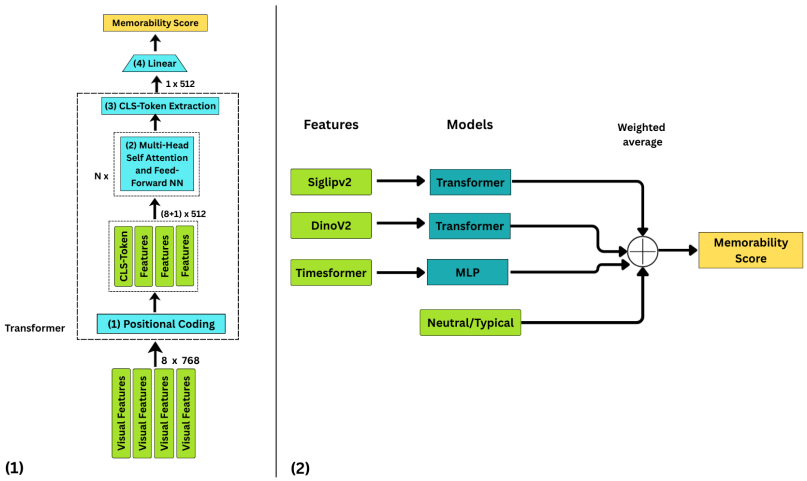

Figure 1: The proposed transformer architecture, alongside the various late fusion strategies proposed for Challenge 1.1.

$\mathbb{R}^{8\times512}$ with a CLS-token appended to the start of the sequence, resulting in $H \in \mathbb{R}^{(8+1)\times512}$, and a standard feed-forward neural network as commonly deployed in transformers [11]: $\mathbb{R}^{(8+1)\times512} \rightarrow \mathbb{R}^{(8+1)\times512}$, (3) CLS-token pooling, which extracts the CLS-token $H_{cls} \in \mathbb{R}^{1\times512}$, and provides that to (4) a sigmoid-activated single-layer perceptron that converts H_{cls} to memorability score m_V, $m_V \in [0, 1]$.

Multi-Layer Perceptron For the motion features, we design a multi-layer perceptron composed of three linear layers, drop-out layers between each linear layer, with two GELU [12] activation functions and a sigmoid activation function in the final layer. This model utilizes the motion features to predict the memorability score m_{mot}.

3. Experimental Design and Results

Challenge 1.1 of MediaEval 2026's [1] Long term memorability track provides the MovieMem [10] dataset. It comprises of 520 videos and memorability scores in the development set and 139 samples in the test set. The "neutral/typical" labels, which designate if a source film is recognizable from the video clip, are also used in our modelling strategies. All of the models used are optimized using the mean square loss function, using the AdamW optimizer. The transformer models were optimized using a learning rate of $5e^{-5}$, while the multi-layer perception was optimized with a learning rate of $5e^{-4}$. We perform 4-fold cross-validation split based on the movie sources, where 75% and 25% are used as training and validation in each fold with no overlap between them. We use the (45^{th}) epoch weights of the transformer models and the (25^{th}) epoch weight of the MLP model to create predictions for each of the late fusion tasks. In-line with the benchmark guidelines [1], we report the Spearman's rank correlation coefficient (SRCC) for the cross validation and test set.

Fusion Strategies The various feature representations, neural network architectures and metadata described in Section 2 are used in late fusion settings, where each model is a weighted sum of each model's prediction. This choice was largely motivated by the effectiveness of late fusion strategies from previous work [2, 3], especially late fusion of visual features. We consider the following five modeling strategies, with the weights determined by a hyper-parameter

search summarized in Table 1:
Run 1 consists of two transformers' predictions, trained separately using SigLip2 and DinoV2 visual features. This run examines the extent to which SigLip2 semantically grounded features complement the DINOv2 features, known for their strong encoding of low-level visual characteristics.
Run 2 consists of one transformer, trained on SigLip2 frame-features, and the $0/1$ neutral and typical labels. This run examines the effectiveness of the neutral labels as a predictive feature, when used in conjunction with our proposed video transformer.
Run 3 consists of the two transformers from Run 1, alongside the neutral and typical labels.
Run 4 consists of one transformer, trained on SigLip2, alongside an MLP trained on the TimeSFormer features. This run demonstrates the effectiveness of the motion and text-grounded visual features.
Run 5 consists of the two transformers from Run 1, alongside an MLP trained on the TimeSFormer features.

Results The results of our experiments (summarized in Table 1) demonstrate the effectiveness of our proposed transformer architectures and of our chosen features. While **Run 1** and **Run 5** performed comparably to the previous editions best approach [2], **Run 4** outperformed their proposed strategy (SRCC: 0.332 vs 0.344 respectively), indicating that our proposed architecture, with visual and motion features is effective. However, **Run 2** and **Run 3** clearly indicate that the presence of neutral features dramatically improves the models performance (SRCC 0.427 and 0.393 respectively), underscoring that contextual features are necessary for state-of-the-art long-term video memorability prediction.

Run Name	Features	Weights	Cross Validation	Test
Run1	SigLip2, Dinov2	0.5, 0.5	0.345± 0.082	0.318
Run2	SigLip, N/T	0.75, 0.25	0.408 ± 0.102	**0.427**
Run3	SigLip2, DinoV2, N/T	0.4, 0.4, 0.2	**0.427** ± 0.089	0.393
Run4	SigLip2, Motion	0.75, 0.25	0.330 ± 0.072	0.344
Run5	SigLip2, Dinov2, Motion	0.4, 0.4, 0.2	0.339 ± 0.076	0.310

Table 1
This table presents the results of our submitted runs. For each run, we report the features used and their corresponding weights in the late-fusion strategy. We use *N/T* to denote the neutral/typical video annotations described in Section 2.1.

4. Discussion and Outlook

As noted in Section 3, combining visual features at multiple levels of granularities with contextual features such as the neutral and typical labels, is highly effective in video memorability prediction. Although these annotations are informative, the specific characteristics that cause a clip to be classified as neutral or typical remain unclear, as no explicit criteria are provided for the labeling process. Consequently, a better understanding of the factors underlying these labels may help identify the contextual elements that contribute to a clip's memorability. We hypothesize that a scene's typicality could be its relevance to the overall narrative of the film. For example, typical scenes may correspond more closely to key events described in a film's synopsis, whereas neutral scenes may be less central to the plot. Future work could investigate this relationship to develop richer contextual representations that could further improve movie-clip memorability prediction.

Acknowledgements

This work is partially supported by project ROBUST: Trustworthy AI-based Systems for Sustainable Growth with project number KICH3.LTP.20.006, which is (partly) financed by the Dutch Research Council (NWO), RTL, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023.

Declaration on Generative AI

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

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: Proc. of the MediaEval 2026 Workshop, Amsterdam, The Netherlands and Online, 2026.

[2] M. Adeel, K. Fatima, M. I. Ayoubi, M. Usmani, M. A. Tahir, Exploring visual, textual, and engagement features for memorability predictions, in: Working Notes Proc. of the MediaEval 2025 Workshop, Dublin, Ireland and Online, In press.

[3] A. Ganesh, I. Huijben, B. Khaertdinov, I. Janssen, M. Popa, N. Tintarev, Dacs-um-rtl: Early fusion and pre-text task learning for video memorability prediction, in: Proc. of the MediaEval 2025 Workshop, Dublin, Ireland and Online, CEUR Workshop Proceedings, In press.

[4] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, et al., Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, arXiv preprint arXiv:2502.14786 (2025).

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).

[6] I. Martín-Fernández, R. Kleinlein, C. Luna-Jiménez, M. Gil-Martín, F. Fernández-Martínez, Video memorability prediction from jointly-learnt semantic and visual features, in: Proceedings of the 20th international conference on content-based multimedia indexing, 2023, pp. 178–182.

[7] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).

[8] Y. Liu, Y. Zhang, D. Ghosh, L. Schmidt, S. Yeung-Levy, Data or language supervision: What makes CLIP better than DINO?, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, Association for Computational Linguistics, Suzhou, China, 2025, pp. 1868–1874. URL: https://aclanthology.org/2025.findings-emnlp.98/. doi:10.18653/v1/2025.findings-emnlp.98.

[9] G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding?, in: Icml, volume 2, 2021, p. 4.

[10] R. Cohendet, K. Yadati, N. Q. K. Duong, C.-H. Demarty, Annotating, understanding, and predicting long-term video memorability, in: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ICMR '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 178–186. URL: https://doi.org/10.1145/3206025.3206056. doi:10.1145/3206025.3206056.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[12] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), 2023. URL: https://arxiv.org/abs/1606.08415. arXiv:1606.08415.