# DACS-UM-RTL: Early Fusion and Pre-text task learning for Video Memorability Prediction

Aashutosh Ganesh[1,*,†], Iris Huijben[1,†], Bulat Khaertdinov[1], Iskaj Janssen[2], Mirela Popa[1] and Nava Tintarev[1]

[1]Department of Advanced Computing Sciences, Maastricht University, The Netherlands
[2]RTL, Hilversum, The Netherlands

Abstract

This edition of MediaEval 2025 introduces two challenges : (1.1) predicting the long term memorability of a video clip and (1.2) classifying whether a person has seen a video based on their EEG recordings. For challenge 1.1, we proposed an early fusion strategy, using a transformer to learn dependencies between the video frames and a machine generated text description. Our approach achieved the highest Spearman's rank correlation coefficient with human annotations (0.259), outperforming both proposed uni-modal models (0.246 for the visual model and 0.170 for the text model) and a late fusion strategy (0.246). For challenge 1.2 we leveraged a 1D-CNN architecture on the raw EEG epochs and experimented with different self-supervised pre-training strategies for the encoder. The best test set AUROC (0.635) was achieved with an end-to-end-trained model without pre-training.

## 1. Introduction

This year's edition of MediaEval 2025 [1] memorability estimation challenge aims to identify properties of movie clips that make them memorable. The challenge uses the MovieMem dataset [2], which provides annotated data capturing two aspects of video memorability. Challenge 1.1 involves predicting, given a movie clip, how likely it is to be recalled upon re-watching. Challenge 1.2 involves predicting whether a participant has previously seen a video based on their EEG recorded during seeing a video, which is formulated as a binary classification task. This paper highlights our contribution to both challenges.

**Challenge 1.1**. Our approach follows a multi-modal fusion paradigm. Prior work mainly employs late fusion [3, 4] or early fusion with linear predictors/RNN [5]. Our best approach utilizes an early fusion strategy which introduces a transformer-like neural network to learn fine-grained relationships between the visual and text features. Our experiments reveal that the early fusion strategy achieves the highest Spearman's rank correlation coefficient (SRCC) when compared to its uni-modal (text and visual) counterparts and a late fusion approach adapted from [4, 6].

**Challenge 1.2**. As opposed to recent works on recall prediction from EEG that leveraged Scalograms or ERPs [7, 8], we leverage a convolutional classifier on the raw EEG epochs, and test several training strategies based on Time Contrastive Learning (TCL) [9] for pre-training the encoder. Our variant, TCL-regress, which uses sliding windowing to enlarge the data

*MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online*
*Corresponding author.
†These authors contributed equally.
Aashutosh.Ganesh@maastrichtuniversity.nl (A. Ganesh); iris.huijben@maastrichtuniversity.nl (I. Huijben); b.khaertdinov@maastrichtuniversity.nl (B. Khaertdinov); Iskaj.Janssen@rtl.nl (I. Janssen); mirela.popa@maastrichtuniversity.nl (M. Popa); n.tintarev@maastrichtuniversity.nl (N. Tintarev)

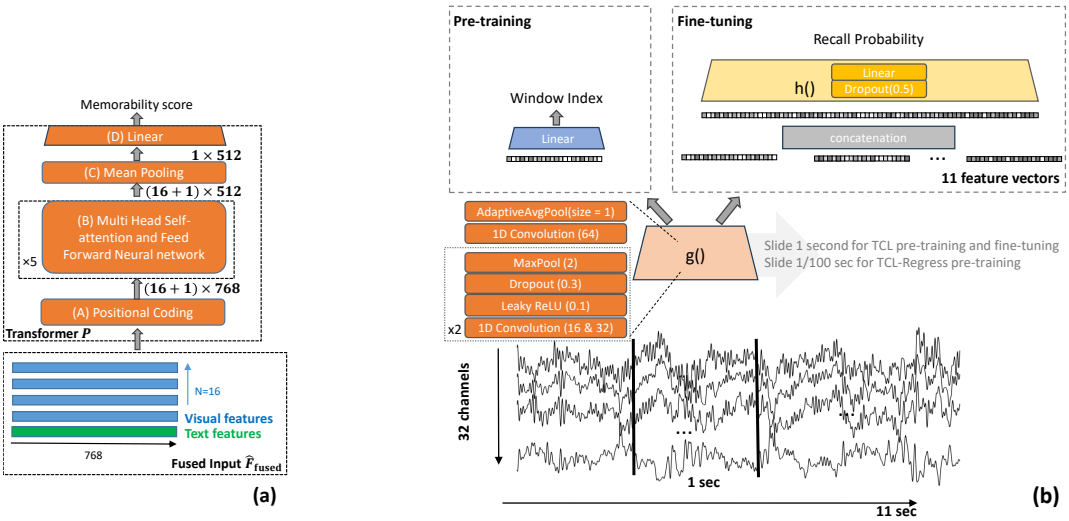


**Figure 1:** An overview of the early fusion strategy proposed for Challenge 1.1 (a) and the pre-training strategy and architecture used in Challenge 1.2 (b).

set, improved performance over no pre-training on 5-fold cross-validation. This result is not reflected in the hold-out test set of the challenge, which achieved comparable results using no pre-training.

## 2. Approach

### 2.1. Challenge 1.1: Video Memorability Regression

**Visual features.** Consider a video $X = \{f_1, f_2, \dots f_N\}$ where $f_i$ denotes the frame at time index $i$ and $N$ is the number of sampled frames. A feature extractor $M_V$ converts each frame to representation $\hat{f}_i$ producing $\hat{F} = M_V(X) = \{\hat{f}_1, \hat{f}_2, \dots \hat{f}_N\}$. For feature extraction, we employ the vision–language model SigLIP [10] (utilizing the default settings provided by Huggingface [11], returning $\mathbb{R}^{1\times768}$ per frame), which demonstrates stronger performance than well-known CLIP [12] in downstream tasks such as text to image retrieval. We uniformly sample $N = 16$ frames per video, resulting in a feature matrix $\hat{F} \in \mathbb{R}^{16\times768}$. A transformer architecture $P$ converts the sequential input $\hat{F}$ to a memorability score $m_V = P(\hat{F}) \in [0, 1]$. The architecture comprises four components: (A) sinusoidal positional encoding [13], (B) 5 stacked transformer blocks, each containing multi-headed self-attention [13] (8 heads with 64 dimensions), $\hat{F}$ is projected to $H \in \mathbb{R}^{16\times512}$, and a standard feed-forward neural network as commonly deployed in transformers [13]: $\mathbb{R}^{16\times512} \rightarrow \mathbb{R}^{16\times512}$, (C) Mean-pooling across the temporal dimension to acquire $H_{agg} \in \mathbb{R}^{1\times512}$, and (D) a sigmoid-activated single-layer perceptron that convert $H_{agg}$ to memorability score $m_V$.

**Text features.** We use SmolVLM2 [14] to generate a caption $T$ per video. A text encoder $M_t$ maps $T$ to feature vector $\hat{T} = M_t(T)$. For $M_t$ we use DistilRoBERTa-v1 [15], as it can create vector embeddings from paragraph-length text, where $M_t(T) \in \mathbb{R}^{1\times768}$. A 4-layer perceptron (with ReLUs and final sigmoid), predicts memorability score $m_T \in [0, 1]$ for $\hat{T}$. The proposed early fusion approach uses SigLIP's[10] text encoder, where $M_t(T) \in \mathbb{R}^{1\times768}$.

Using these visual and textual features, we introduce our proposed early fusion strategy and we describe a late fusion strategy adapted from [3]. **(1) Early Fusion** - text embedding $\hat{T}$ is concatenated to the visual frame-wise features leading to: $\hat{F}_{\text{fused}} = \{\hat{f}_1, \hat{f}_2, \dots, \hat{f}_N, \hat{T}\}$ . $\hat{F}_{\text{fused}} \in \mathbb{R}^{(16+1)\times768}$ is provided as input to transformer model $P$ as visualized by Figure 1a).

This approach to early fusion should enable the multi-head self-attention to learn relationships between the complete video's text semantics and the individual visual frame features. **(2) Late Fusion:** We compute the weighted average between the predictions from visual features and text features: $m_{\text{fused}} = \alpha\, m_{\text{V}} + (1-\alpha)\, m_{\text{T}}$, where $\alpha \in [0, 1]$ is a fixed fusion weight.

### 2.2. Challenge 1.2: Video Recall Classification from EEG

We process all EEG epochs by downsampling the signals to 100 Hz and applying z-score normalization per subject, per epoch, and per channel. We denote the processed epochs per subject $k$ as $\mathbf{X}^{(k)} \in \mathbb{R}^{E\times C\times N}$, with $E$ being the number of videos seen/epochs recorded, $C = 32$ the number of EEG channels, and $N = Tf_s$, the number of time stamps, with $T$ the total recording time in seconds, capped to 11 s, and $f_s$ the sampling frequency in Hz. We denote the $j^{\text{th}}$ 1 s window within epoch $i$ as: $\mathbf{X}_i^{(k)}[j] \in \mathbb{R}^{C\times f_s}$. The binary label (recognized or not) of epoch $i$ is denoted with $\mathbf{y}_i^{(k)}$. We define neural network $f_\theta$ that comprises two parts. The first part is a 1D convolutional encoder $g$ that embeds a window to: $\mathbf{z}_i^{(k)}[j] = g(\mathbf{X}_i^{(k)}[j]) \in \mathbb{R}^{D'}$. The full embedding of epoch $i$ is the concatenation of all window embeddings, resulting in $\mathbf{z}_i^{(k)} \in \mathbb{R}^{D}$, with $D = 11D'$ . The second part is a classifier $h$, that applies dropout ($p = 0.5$), followed by a sigmoid-activated linear layer to predict the probability that epoch $i$ was recognized: $\hat{\mathbf{y}}_i^{(k)} = h(\mathbf{z}_i^{(k)}) \in [0, 1]$. Figure 1b shows the architecture, while the training strategies are detailed in Sec. 3.2.

## 3. Experimental Design and Results

### 3.1. Challenge 1.1

The MovieMem dataset [2] comprises 520 videos in the development set and 139 in the test set. Each model is trained by minimizing the mean squared loss, using the Adam optimizer, with a learning rate of $5e^{-4}$ and cosine scheduling. We perform 5-fold cross-validation, where 80% and 20% are using as training and validation in each fold with no overlap between them. We report the average SRCC as the mean over each fold (reported as 'cross val'). For the hold-out test set, models are trained on the full development set, and we report test set results from the final ($25^{th}$) epoch of each model.

We evaluate several configurations: *'Uni-Modal Vision'*: using $m_V$, *'Uni-Modal Text'*: using $m_T$, *'Early Fusion'*, *'Late fusion'* (with $\alpha = 0.7$, chosen based on best cross-validation SRCC), and Bayesian Ridge Regression (*'BRR'*; with default settings from [16]) trained on visual features $\hat{F}$. As shown in Table 1, the early fusion strategy achieves the best SRCC on the test set (0.259) and the second-best result in cross-validation (0.368). Despite the discrepancy between validation and test performance, both early and late fusion consistently outperform their unimodal counterparts.

**Table 1:** Spearman' rank correlation coefficients (SRCC) for video memorability prediction (1.1).

| Model Strategy | Cross-val | Test set |
|---|---|---|
| Uni-modal Vision | 0.362 ± 0.10 | 0.224 |
| Uni-modal Text | 0.301±0.03 | 0.160 |
| Early Fusion | 0.368 ± 0.09 | **0.259** |
| Late Fusion | **0.378**± 0.06 | 0.246 |
| BRR | 0.243± 0.07 | 0.246 |

**Table 2:** AUROC values for different training strategies in the recall prediction from EEG (1.2).

| Training strategy | Cross-val | Test set |
|---|---|---|
| No pre-training | 0.628 ±0.015 | **0.635** |
| TCL | 0.614± 0.010 | 0.615 |
| TCL-regress | **0.634**±0.024 | 0.630 |
| TCL-regress ensemble | NA | 0.612 |
| TCL-regress exp-sampling | **0.634**±0.016 | 0.618 |

### 3.2. Challenge 1.2

We experiment with four strategies to train the classification model. First, we run end-to-end training with the standard cross-entropy loss ('*no pre-training*'). Second, we pre-train the encoder using the self-supervised learning strategy TCL [9]. This trains the encoder and a classifier to predict window index $j$ for each non-overlapping 1-sec window $\mathbf{X}_i^k[j]$. After pre-training, we remove the classifier to run end-to-end finetuning of the full model with cross-entropy loss ('*TCL*'). Third, we use overlapping 1-sec windows, and use their relative position within the epoch as the label for the TCL task, resulting in pre-text label between $[0, 1]$. This increases the training data and reformulates TCL into a regression problem, so mean squared error is minimized instead. End-to-end finetuning of the complete model ('*TCL-regress*') is done subsequently. Fourth, we hypothesize that earlier windows in the epoch may be of higher importance to the final classification task, as they are closer to the moment of video onset. Therefore, we test a strategy ('*TCL-regress exp-sampling*') where earlier 1-sec windows are more often sampled from the epoch according to an exponentially-decreasing probability function with decay factor 500. The proposed strategies are trained on the same seed, with 5-fold cross-validation on the *devset* and once again on the full *devset*. Additionally, we use the 5 cross-validation *'TCL-regress'* models to create an ensemble prediction on the test set by averaging their predicted probabilities, referred to as *'TCL-regress ensemble'*. We report AUROC average over the cross-validation sets, and on the test set. We used Adam optimizer with a base learning rate of 1e-4 for pre-training, and 5e-5 for full model fine-tuning/training. The learning rate reduced a factor 10 every 20 epochs of no validation loss improvement (or training loss in case of full devset training). Table 2 shows that *TCL-regress* achieves the highest AUROC during cross-validation, while *'no pre-training'* performs best on the test set. The '*TCL-regress exp-sampling*' strategy did not show improvement over '*TCL-regress*' for either the cross-validation or the test set and the ensembling strategy neither improved test set results.

## 4. Conclusions

**In Challenge 1.1**, all models (except BRR) performed better in cross-validation than on the test set, with substantial variance in SRCC scores. The lower test-set performance likely reflects a degree of overfitting to the development data and the heterogeneity of the movie domain, as each film's unique visual style and narrative can challenge generalization. The disparity between the late-fusion cross-validation and test-set results could also be attributed to miscalibration. Since late fusion relies on an ensemble strategy, its performance can be undermined when the component models are miscalibrated [17]. Improving calibration, and using captioners that incorporate richer details, such as film elements, may enhance memorability prediction and yield new insights into what makes movie shots memorable.

**In Challenge 1.2**, the TCL regression setting surpassed "no-training" in cross-validation, suggesting that either the additional data generated by TCL-regress or the pretext task itself contributed positively. However TCL-regress did not improve hold-out test set performance. During cross-validation we observed overfitting despite high dropout. This could explain the discrepancy between cross-validation and test set results, as the models for the latter could not leverage early-stopping due to the lack of a validation set. The best performing model on the test set was the end-to-end trained model (without pre-training), which achieved an AUROC of 0.635. Lastly, the model was subject-independent, while inter-personal differences are known to be large in EEG. Subject-conditioning [8] might, therefore, improve future model's performance, at the cost of flexibility to unseen subjects.

## Declaration on Generative AI

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

## 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. 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.

[3] M. M. A. Usmani, H. Faisal, M. A. Tahir, The impact of transformers ensemble on model memorability and generalizability., in: MediaEval, 2023.

[4] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, A. Oliva, Multimodal memorability: Modeling effects of semantics and decay on video memorability, in: European Conference on Computer Vision, Springer, 2020, pp. 223–240.

[5] A. Viola, S. Yoon, A hybrid approach for video memorability prediction., in: MediaEval, 2019.

[6] L. Sweeney, G. Healy, A. F. Smeaton, Predicting media memorability: comparing visual, textual and auditory features, arXiv preprint arXiv:2112.07969 (2021).

[7] L. Sweeney, A. Smeaton, G. Healy, Memories in the making: Predicting video memorability with encoding phase eeg, in: Proceedings of the 20th International Conference on Content-Based Multimedia Indexing, 2023, pp. 183–187.

[8] 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., 2023.

[9] A. Hyvarinen, H. Morioka, Unsupervised feature extraction by time-contrastive learning and nonlinear ica, Advances in neural information processing systems 29 (2016).

[10] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 11975–11986.

[11] 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).

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[13] 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).

[14] A. Marafioti, O. Zohar, M. Farré, M. Noyan, E. Bakouch, P. Cuenca, C. Zakka, L. B. Allal, A. Lozhkov, N. Tazi, V. Srivastav, J. Lochner, H. Larcher, M. Morlon, L. Tunstall, L. von Werra, T. Wolf, Smolvlm: Redefining small and efficient multimodal models, 2025. URL: https://arxiv.org/abs/2504.05299. arXiv:2504.05299.

[15] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[16] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.

[17] A. Kumar, T. Ma, P. Liang, A. Raghunathan, Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift, in: Uncertainty in Artificial Intelligence, PMLR, 2022, pp. 1041–1051.