International Journal of Computer Vision (2026) 134:298
https://doi.org/10.1007/s11263-026-02880-6

A Review of Computational Memorability: A Benchmark Framework

Mihai Gabriel Constantin[1] · Claire-Hélène Demarty[2] · Camilo Fosco[3] · Sebastian Halder[4] · Graham Healy[5] · Bogdan Ionescu[1] · Stefan Valentin Luncanu[1] · Iván Martín-Fernández[6] · Ana Matran-Fernandez[4] · Rukiye Savran Kiziltepe[7] · Alan F. Smeaton[5] · Liviu-Daniel Stefan[1] · Lorin Sweeney[5] · Alba García Seco de Herrera[8]

Received: 9 September 2025 / Accepted: 5 May 2026
© The Author(s) 2026

Abstract

One of the powers of visual media lies in its ability to create a lasting impression on the viewer's memory. In this digital age, where media is abundant and attention spans are fleeting, the task of predicting which content will stick in the viewer's mind has become a critical challenge in computer vision. Computational memorability seeks to address this by developing models that estimate how memorable a piece of media is likely to be. In this review we focus on the MediaEval Predicting Video Memorability benchmark, a recurring evaluation task that has run annually since 2018. This benchmark provides a unique and consistent framework for researchers to compare and refine their memorability prediction techniques using standardised datasets and metrics. Its reproducible framework has proven invaluable for tracking progress and fostering innovation in this rapidly evolving domain. We analyse the evolution of the benchmark across its 2018–2023 editions, discussing the challenges that still remain, such as the need for more interpretability in models and the difficulty of predicting subjective and context-dependent memorability. By analysing and synthesising the collective insights gained from this task, we endeavour to inspire new avenues of inquiry and drive progress towards a more comprehensive understanding of this topic.

Keywords Memorability · Multimedia · Machine Learning · Deep Learning

Communicated by Yasuyuki Matsushita.

Mihai Gabriel Constantin, Claire-Hélène Demarty, Camilo Fosco, Sebastian Halder, Graham Healy, Bogdan Ionescu, Stefan Valentin Luncanu, Iván Martín-Fernández, Ana Matran-Fernandez, Rukiye Savran Kiziltepe, Alan F. Smeaton, Liviu-Daniel Stefan, Lorin Sweeney, and Alba G. Seco de Herrera contributed equally to this work.

✉ Alba García Seco de Herrera
alba.garcia@lsi.uned.es

Mihai Gabriel Constantin
mihai.constantin84@upb.ro

Claire-Hélène Demarty
claire-helene.demarty@interdigital.com

Camilo Fosco
camilofosco@gmail.com

Sebastian Halder
s.halder@essex.ac.uk

Graham Healy
graham.healy@dcu.ie

Bogdan Ionescu
bogdan.ionescu@upb.ro

Stefan Valentin Luncanu
stefan.luncanu@stud.fsa.upb.ro

Iván Martín-Fernández
ivan.martinf@upm.es

Ana Matran-Fernandez
amatra@essex.ac.uk

Rukiye Savran Kiziltepe
rukiyekiziltepe@ankara.edu.tr

Alan F. Smeaton
alan.smeaton@dcu.ie

Liviu-Daniel Stefan
liviu_daniel.stefan@upb.ro

Lorin Sweeney
lorin.sweeney@dcu.ie

[1] AI Multimedia Lab, CAMPUS Research Institute, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania

[2] R&I France, InterDigital, Paris, France

[3] CSAIL, MIT, Cambridge, USA

Published online: 30 May 2026

1 Introduction

Computational memorability is an emerging field at the intersection of cognitive science and artificial intelligence that seeks to model and predict how likely a piece of content–such as an image or video–is to be remembered by viewers. Understanding memorability has broad implications for content creation, education, advertising, and accessibility, as it enables the design of media that is more engaging and impactful. In recent years, advances in machine learning and deep learning have enabled researchers to develop models that can estimate memorability scores directly from visual and audio features. These models are trained on annotated datasets where human participants indicate whether they remember previously seen content, allowing for the creation of predictive systems that generalise across different types of media. Within this context, benchmarking initiatives such as The Benchmarking Initiative for Multimedia Evaluation (MediaEval)[1] have played a crucial role in standardising evaluation and fostering progress in the field.

Research in neuroscience and psychology has long pointed to the structured memorability of multimedia content, from demonstrating the remarkable capacity of human visual memory (Shepard, 1967; Standing, 1973), to showing its consistency across observers, even when intertwined with subjective judgments (Jaegle et al., 2019; Isola et al., 2011), largely driven by scene statistics and object-level properties (Konkle et al., 2010, b), and culminating in recent studies confirming that memorability reflects intrinsic properties of stimuli rather than individual preference alone (Lin et al., 2021; Xie et al., 2020). Historically, several attempts have been made to create datasets and conduct studies on human participants and assessors to measure either short- or long-term memorability. Gygli et al. (2013) presents a foundational study centred on interestingness and its covariates, including assumed memorability (human judgment of whether images will be remembered) and actual memorability (whether participants truely remember the images, measured through actual memory tests). While analysing various correlations between different concepts related to human perception of visual media, the authors found that memorability is inversely correlated with interestingness and, in line with the conclusions of Isola et al. (2011), that assumed memorability is negatively correlated with actual memorability. They concluded that "What a human observer finds interesting is what he wants to remember and believes he will. Unfortunately, the latter is often not the case". Another important study (Isola et al., 2011) created a dataset composed of 2222 target images measured for memorability. The authors studied a range of features and concepts that correlated with memorability in everyday images, considering both objects in the images and global and local visual attributes. A more modern approach to the study of memorability has been presented by Khosla et al. (2015), where the LaMem dataset was constructed. The authors sampled images from a variety of sources related to social media, image popularity, affective content, abnormal object presence, and image saliency datasets, and observed that some types of content, such as image popularity and abnormal objects, tend to be more memorable. Other widely used datasets are discussed later in this paper, including VideoMem (Cohendet et al., 2019) and Memento10k (Newman et al., 2020), as they are used in building various editions of our proposed benchmarking tasks.

MediaEval is a multi-task collaborative benchmarking event for multimedia retrieval, access, exploration and analysis. Since 2018, the "Predicting Video Memorability" task has been organised annually as part of MediaEval, aiming to understand what makes a video more, or less, memorable. Several articles describe the overall evolution of the set of subtasks within the MediaEval benchmark and the various approaches that were used to create the resources used to support them Gabriel Constantin et al. (2024); Sweeney et al. (2023, 2022); Kiziltepe et al. (2022); De Herrera et al. (2021); Constantin et al. (2019); Cohendet et al. (2018). Throughout the different editions of the Predicting Video Memorability task, we focused on exploring several aspects of memorability, and of approaches that seek to understand and predict the memorability of short-form videos. As we will show in this paper, we proposed the study of short- and long-term memorability, integrated different open datasets into our tasks, analysed the generalisation power of memorability prediction systems, and looked into electroencephalogram (EEG) as possible indicators and features for memorability prediction.

The Predicting Video Memorability task has been a consistent task of MediaEval, with annual editions running from 2018 through 2023 inclusive, and a new iteration finished in October 2025 (as MediaEval took a hiatus in 2024). The 2018–2023 period represents a cohesive phase of the task, focused on the memorability of short-form, social media-style videos. This phase enabled the development and benchmarking of models under controlled yet diverse con-

4 School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom

5 School of Computing, Dublin City University, Dublin, Ireland

6 Grupo de Tecnología del Habla y Aprendizaje Automático (THAU Group), Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid (UPM), Madrid, Spain

7 Department of Software Engineering, Ankara University, Ankara, Türkiye

8 ETSI Informática, UNED, Madrid, Spain

[1] https://multimediaeval.github.io/

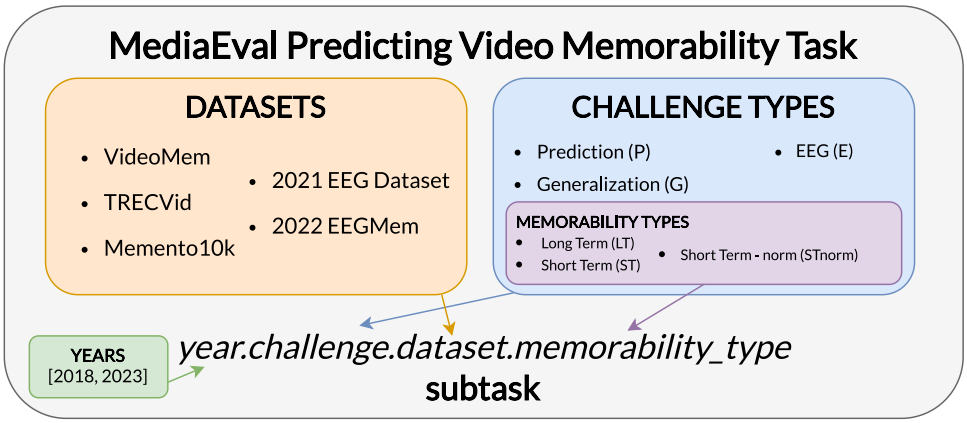


Fig. 1 Structure of the MediaEval Predicting Video Memorability Task. Each year, a series of subtasks is introduced, each presenting a distinct challenge that makes use of a specific dataset and memorability type.

ditions tailored to short video content. In contrast, *the 2025 edition marks a significant shift in the task's scope*, expanding it to include longer videos, notably movie excerpts and commercial content. We therefore consider *2023 to be a natural cutoff point*, and in this paper, we propose an in-depth analysis, examining the benchmarking framework, datasets used throughout the editions of the memorability task, methods and modalities employed by the participants, and their results across the 2018–2023 editions by participants in the task. We also perform experiments with ensemble techniques, fusing the results sent by the participants in order to measure a "maximal" performance level given the systems proposed thus far. Finally, we aggregate all the findings and correlations and propose a set of high-level lessons that can be drawn from the experience and analysis of the 2018–2023 editions of the task, as well as look into some open research questions that still remain unanswered. Because this article is conceived as a retrospective analysis of the MediaEval Predicting Video Memorability task, our scope is deliberately limited to the 2018–2023 cycle, which constitutes a methodologically coherent phase of the benchmark. Accordingly, the conclusions presented in this work should be interpreted within the context and constraints of this specific benchmark framework.

To ensure consistency and readability throughout the paper, we introduce a terminology that structures the key concepts involved in organising the challenge. This structure is illustrated in Figure 1 and includes the following terms:

- *Task* refers to the overall Predicting Video Memorability benchmarking competition, without looking into a specific edition or partition.
- *Viewers* or *annotators* refer to people that were employed during the annotation process of the datasets we used, viewing the videos and signaling whether they recall seeing the videos or not, thus helping us in creating the ground-truth memorability values associated with each dataset.
- *Participants* and *teams* refer to the teams that have participated in the various editions of the MediaEval benchmarking task by submitting their runs and submitting a paper to the MediaEval proceedings that describes their runs.
- *Run* or *submission* denominates a single method (and the corresponding set of predictions) created by one of the teams and submitted to the organisers, that attempts to solve one subtask for an edition of the MediaEval Predicting Video Memorability task.
- *Dataset* denominates one of the five open datasets used throughout the different editions of the Predicting Video Memorability task.
- *Memorability type*, as used in our scenarios, nominates one of the two main types of memorability: short-term memorability, denoting the ability to recall a video in a short time after initial viewing, and long-term memorability, referring to the ability to recall a video days after initial viewing.
- *Challenge type* defines one of the three types of challenges we proposed to participants: video-based *prediction* challenges, where memorability must be predicted using approaches that process video data and associated metadata, video-based *generalisation* challenges, where participants must train on one dataset and test on another, thus validating the generalisation capabilities of their approaches, and *EEG-based* prediction.
- *Subtask* defines the numerous challenges that we propose for participants throughout the editions of the memorability task. Each subtask is defined by a year when the challenge was proposed, the dataset used in the subtask, the challenge type and the memorability type.

The rest of the paper is structured as follows: The datasets used and the subtasks in the Predicting Media Memorability task have evolved over the years and are described in Sections 2–3. An overview of the main techniques that achieved best results in the benchmark is given in Section 4.

A set of ensemble approaches based on applying late fusion approaches to all the runs submitted by participants throughout the years is discussed in Section 5. The main lessons learned are described in Section 6 and conclusions are given in Section 7.

2 Subtasks and Participation Over the Years

Table 1 provides an overview of the subtasks featured in each edition of the Predicting Video Memorability benchmark. The set of core *video-based prediction* subtasks challenged participants to automatically estimate memorability scores for videos – that is, the likelihood that a video will be recalled upon subsequent viewings, either in the *short term* or the *long term*. The set of *generalisation* subtasks evaluated the robustness of prediction models across different datasets and data sources: participants are required to train their systems on one dataset and evaluate them on a different dataset, thereby testing their ability to generalise to new video sources. Finally, the set of *electroencephalogram (EEG)-based prediction* subtasks introduced a neurophysiological dimension by incorporating EEG signals recorded from viewers as they watched the videos. The goal was to predict whether a viewer would remember a video or not, either using only EEG data or in combination with features extracted from the videos.

The datasets used during the 2018–2023 editions of the memorability task are characterised by their short-form nature, with durations typically ranging from 5 to 10 seconds. The working model in our use case for memorability is that the viewers of the videos have no personal or emotional connection with those videos. That is, the videos do not show their family, friends or familiar personal context. While the content may include well-known public figures or widely recognisable scenes, it is not expected to involve individuals or places with whom the viewers have a direct, personal relationship. Thus, this task focuses on the kind of rapid, attention-driven content commonly encountered on social media platforms. During the different editions, task participants were provided with extensive datasets of videos, each with memorability annotations, related information, pre-extracted state-of-the-art visual and audio features, and EEG recordings (see Section 3). Furthermore, for some editions, the same set of videos and ground-truth scores were used, thus helping us in understanding the evolution of proposed systems and methods, as well as helping participants in building their systems upon those submitted to previous editions. A breakdown of the datasets and which editions they were used in is presented in Section 3.

Over the 6 completed editions of running the memorability task, 47 teams successfully participated (with submitted runs and papers), with a total of 376 runs submitted for evaluation,

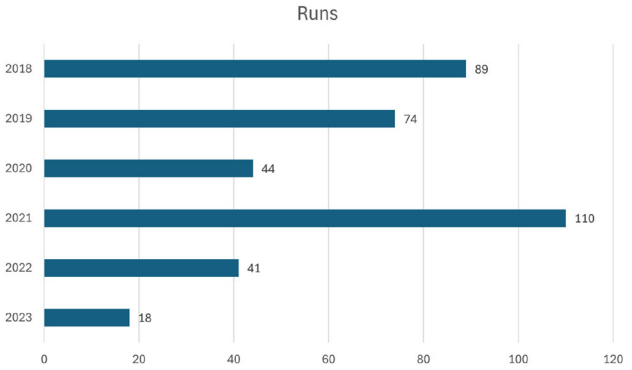


Fig. 2 Annual participation in the MediaEval Predicting Video Memorability task for the 2018–2023 editions, expressed as number of runs submitted per year.

as shown in Figure 2. With the larger availability of benchmarks and datasets nowadays, the task remains attractive for participants. This shows the impact that building data collections and benchmarking tasks and sharing these resources with other researchers can have on research challenges like predicting memorability. The high participation across editions reflects the enduring relevance of understanding and predicting memorability, as well as the fact that memorability prediction is still an open area of research.

3 Data Over the Years

This section provides an overview of the evolution of the datasets used in our proposed task. Over the years, the task has introduced new datasets and refined existing ones to better capture the nuances of video memorability, including variations in video length, content type, and annotation protocols. These changes reflect the growing complexity and ambition of the task, as well as the community's efforts to improve the robustness and generalisability of memorability prediction models. Following this overview, we describe the methodologies employed to compute memorability scores, including the experimental protocols used to collect human annotations and the statistical techniques applied to derive consistent and comparable memorability values across different datasets.

3.1 Datasets

Between 2018 and 2023, three video datasets and two EEG-based datasets have been released as part of the MediaEval Predicting Media Memorability task, each accompanied by pre-extracted features. While the 2025 and 2026 editions introduce new datasets, these fall outside the scope of this paper and are not included in our analysis, as explained in Section 1. Details regarding evaluation protocols and memo-

Table 1 Summary of challenges proposed in the Predicting Video Memorability task across MediaEval editions from 2018 to 2025. Note that MediaEval took a hiatus in 2024, and the 2025 edition finished in October 2025.

	2018	2019	2020	2021	2022	2023	2025 (just finished)
Video-based short-term prediction	✓	✓	✓	✓	✓		
Video-based long-term prediction	✓	✓	✓	✓			✓
Generalisation				✓	✓	✓	
EEG-based memorability prediction				✓	✓	✓	✓

rability score computation are presented in Sections 3.2 and 3.3, respectively. We also note that, for all datasets except Memento10K, annotators viewed the videos together with their corresponding audio during the annotation process. Although Memento10K contains audio tracks, these were not presented to annotators when collecting memorability annotations.

VideoMem VideoMem (Cohendet et al., 2019) is composed of 10,000 7-second long soundless videos shared under a license allowing their use and redistribution in the context of MediaEval. They were extracted from high quality cinematic raw stock footage used by professionals when creating content. They are varied and contain different scenes types. Each video is accompanied with short-term (i.e., measured a few minutes after visualisation) and long-term (i.e., measured 24–72 hours later) memorability scores and a caption that contains a short description of its content. This dataset was used in the 2018, 2019, 2022, and 2023 editions of the task. In both the 2018 and 2019 editions, the same dataset was used, with a consistent split of 80% for training and development and 20% for testing. This dataset served as the single and primary resource of evaluation in both years. In the 2022 edition, the dataset used was the same as in previous editions, with an additional 5% of labeled videos released. The data was split into 70% for training, 15% for validation, and 15% for testing, offering a slightly larger labeled set while maintaining continuity with earlier editions. In 2023, the dataset was only used to evaluate the systems that were submitted, in the generalisation challenge type. In this case, the teams were not allowed to use any VideoMem data for training. Only the 2,000 videos in the testing set (without labels) were released for the teams to perform inference and submit their runs for models trained with other datasets. Figure 3 shows a sample of frames from some of the videos in this dataset and their corresponding memorability scores.

TRECVid 2019 Video-to-Text The TRECVid dataset (Awad et al., 2019) contains 6,000 Vine videos provided under Creative Commons licenses that allow their redistribution. A subset of this dataset was selected and annotated to be used in the 2020 and 2021 MediaEval Predicting Video Memorability editions. The data were divided into training, validation, and testing sets. In 2020, the released data included 590 training and 410 validation videos while in 2021 the validation set was extended up to 1116 videos. In both years, the test set remained unchanged with 500 samples. The TRECVid videos have different audio and visual features in scenes where various actions are performed by different objects. Therefore, these movies include more dynamic actions and audio compared to those in VideoMem. Figure 4 shows a sample of frames from some of the videos in this dataset and their corresponding memorability scores.

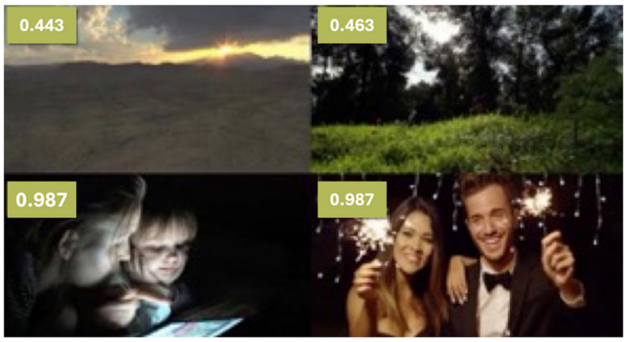


Fig. 3 A sample of frames (and corresponding memorability scores) from videos in the VideoMem dataset.

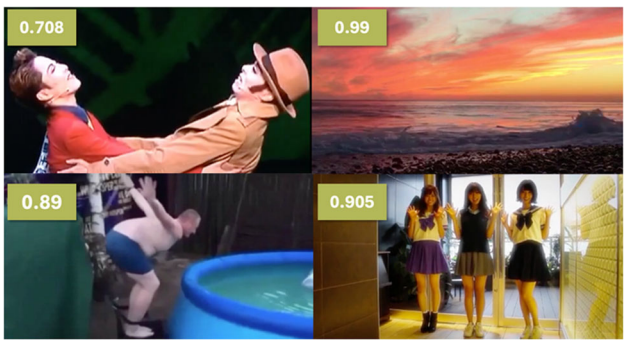


Fig. 4 A sample of frames (and corresponding memorability scores) from videos in the TRECVid dataset.

Memento10K The Memento10K dataset (Newman et al., 2020) contains 10,000 three-second long videos depicting in-the-wild scenes. Each video is annotated with short-term memorability scores, memorability decay values, action labels, and five accompanying captions describing the content of the video. For the purposes of the MediaEval Predict-

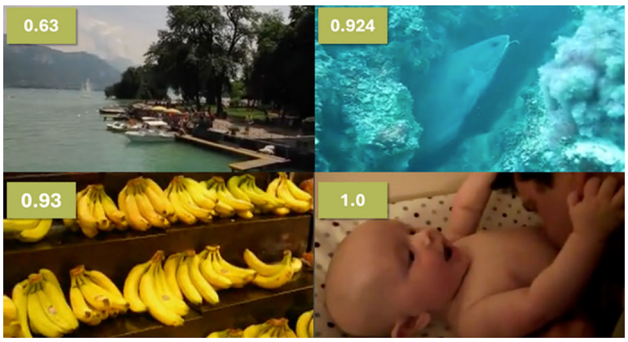


Fig. 5 A sample of frames (and corresponding memorability scores) from videos in the Memento10K dataset.

ing Video Memorability task, 7,000 videos were released for training, 1,500 for validation and the remaining 1,500 videos were used as the test set for scoring submissions. This dataset was used consistently across the 2021, 2022, and 2023 editions of the task. Figure 5 shows a sample of frames from some of the videos in this dataset and their corresponding memorability scores.

2021 EEG dataset The 2021 EEG dataset (Sweeney et al., 2022) was collected as part of a pilot study to check the feasibility of finding neural correlates of memory encoding in a short-term memorability task. The stimuli consisted of 450 videos selected from the TRECVid dataset, with a balanced representation of high, low, and medium memorability scores, ensuring consistency with prior editions of the task. EEG data were recorded from 11 subjects across two locations using a shared experimental protocol. Annotators viewed the videos in a continuous session without distinct encoding and recognition phases. Each video was followed by a prompt asking whether it had been seen before, allowing for binary annotation of short-term memorability. However, the resulting dataset was highly imbalanced, with the majority of the video clips being correctly remembered by the annotators. This imbalance limited the scope of meaningful analyses, particularly in distinguishing neural patterns associated with remembered vs. forgotten content. As a result, the experimental protocol was refined for the 2022 edition of the EEG task to improve data balance and enhance the reliability of subsequent analyses.

2022 EEGMem The 2022 EEG dataset (Sweeney et al., 2023) built on the 2021 pilot study with a revised protocol designed to improve data balance and support long-term memorability analysis. EEG recordings were collected from 12 subjects captured while they watched a subset of the Memento10k (Newman et al., 2020) videos. To assess long-term memory retention, the annotators rewatched the same videos (plus a random selection of videos that had not been shown to them during the encoding phase of the study) through a custom-built online portal between 24–72 hours after a video-EEG recording session, where they were required to indicate for each video whether or not they recognised it. This provided binary ground truth annotations of memory encoding for long-term memorability. As part of the EEG subtask, participants were required to predict whether a given viewer would remember a specific video based on their EEG signals. A cross-subject/transfer learning component was introduced, with data from three annotators that were left out of the training set used to evaluate model transferability across individuals.

3.2 Measuring Video Memorability

Inspired by the works of Isola et al. (2013, 2011), the datasets developed or integrated in the Predicting Video Memorability task use a two-stage annotation process. An example of this process applied to the VideoMem dataset is presented in Figure 6, while some statistics and measurements and particularities of the annotation process for each dataset are presented in Table 2.

The first stage deals with annotating videos for short-term memorability ground-truth scores. Viewers watch a series of videos, containing "target videos" (videos that are part of the final dataset) and "vigilance videos" (videos that are repeated quickly after their occurrence, used to measure viewer attention). While target videos are shown again after tens of other videos (e.g., for the VideoMem collection, they are repeated after 45–100 videos, accounting for only a few minutes after retention), vigilance videos are repeated after only a few other videos (e.g., the repetition of the vigilance videos occurs after 3–6 videos for the VideoMem dataset). The second stage deals with long-term memorability annotation, where after a longer period of time (i.e., 24–72 hours), viewers participate in another round of memorability annotations.

For the VideoMem dataset, the memorability scores were computed from 9,402 viewers for short-term, and 3,246 viewers for long-term memorability. On average, 38 (at least 30) annotations were collected for each video in the short-term task, whereas in the long-term task, a mean of 13 annotations (at least 9) were collected. For the TRECVid 2019 Video-to-Text dataset, a minimum of 14 annotations per video in the short-term memorability step were collected with a mean of 22. For long-term memorability a minimum of 3 and an average of 7 annotations/video were collected. For the Memento10K only short-term memorability annotations were acquired. The memorability scores were computed with 90 annotations/video on average.

Annotators' agreement was computed as defined in Cohendet et al. (2019) and an annotation consistency value is computed for each number N of annotations per video. First only videos with at least N annotations each are considered. This video set is then randomly split in two equal parts, before computing the Spearman rank correlation between

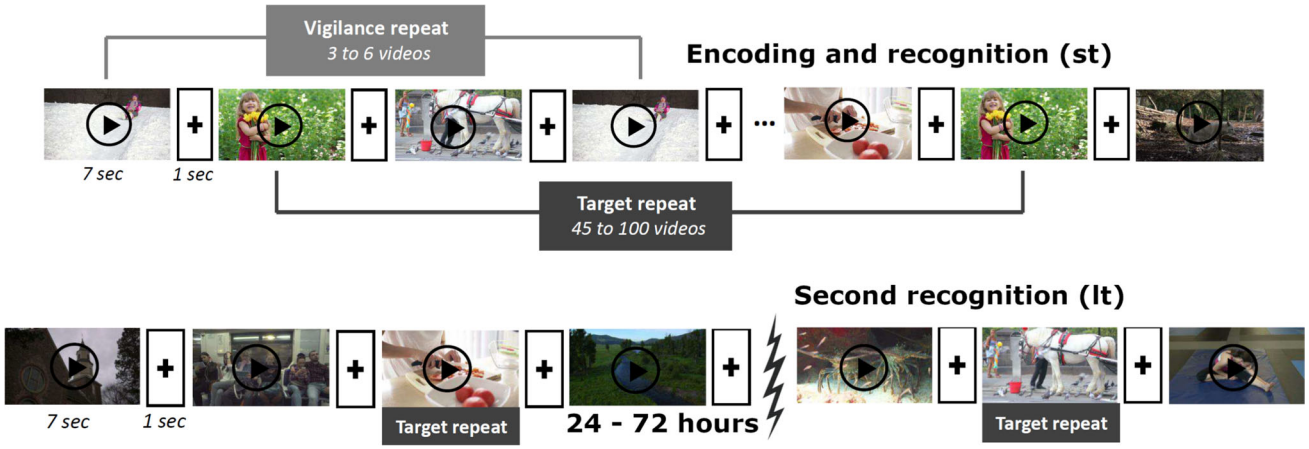


Fig. 6 A graphical representation of the annotation protocol, as presented in Cohendet et al. (2019), showing the differences between short-term (st) and long-term (lt) annotations.

Table 2 Summary of the three visual datasets used in the 2018–2023 editions (VideoMem, TRECVid, Memento10k) of the Predicting Video Memorability task. For this table, *st* denotes short-term annotations, while *lt* denotes long-term.

	VideoMem	TRECVid	Memento10k
Total number of videos	10,000	6,000	10,000
Average duration	7 sec.	6 sec.	3 sec.
Memorability type	short- and long-term	short- and long-term	short-term
Avg. number of annotators per video (st)	38	22	90
Avg. number of annotators per video (lt)	13	7	-
Min. number of annotators per video (st)	30	14	64
Min. number of annotators per video (lt)	9	3	-
Vigilance repeat	3–6 videos	3–6 videos	2–3 videos
Target repeat (st)	45–100 videos	45–100 videos	9–200 videos
Wait time until second stage of annotation (lt)	24–72 hours	24–72 hours	-
Number of videos per viewer in a session (st)	180	180	200
Number of videos per viewer in a session (lt)	120	120	-

the two sub-sets. Computed over VideoMem, the annotation consistency reaches a value of 0.616 for 38 annotations per video, for short-term memory, and 0.364 for 13 annotations per video, for long-term memory. From Cohendet et al. (2019), it can be extrapolated to reach a maximal value of 0.81 for 72 annotations per video, for short-term memory. On Memento10K, it reaches 0.73, for short-term memory and 90 annotations per video. For the TRECVid dataset, inter-annotator agreement was estimated using a logarithmic extrapolation model fitted on merged subsets of the data, following the method used for VideoMem dataset. To ensure reliability, the fitting was performed only on data points where the number of annotators per video was $\leq$ 12, in the case of short-term memory, due to high variability and limited video availability at higher annotation levels. Based on this model, the extrapolated Spearman correlation at 50 annotations per video is approximately 0.1901 for short-term memory. For long-term memory, only a few points have more than 12 annotators per video, leading to the conclusion that our method to extrapolate the Spearman correlation at 50 annotations would not be robust enough. As a conclusion, this and the value of 0.1901 for short-term memory highlight the limited annotation consistency in TRECVid compared to datasets like VideoMem and Memento10K.

Furthermore, as described by Cohendet et al. (2019), in addition to the inclusion of vigilance videos and filtering mechanisms, annotator consistency is ensured through the computation of control recognition rates, which allow for the identification and removal of annotators with abnormal recognition behavior. These rates are established based on in-house tests and define the following thresholds: (i) a minimum correct recognition rate of 15% for long-term annotations; (ii) a maximum false recognition rate of 30% for short-term annotations and 40% for long-term annotations; and (iii) a false recognition rate lower than the true recognition rate for long-term annotations. While these mechanisms

strengthen internal consistency and attention quality across annotators, it is important to note that no measures were implemented to control for potential cultural or demographic bias during dataset creation, as annotator background information was neither collected nor used as part of the filtering process.

3.3 Memorability Scores Over the Years

In the first four editions of the task, each memorability score was calculated as the percentage of correct recognitions among viewers (termed *raw scores*). In the 2018 and 2019 editions, a linear correction was applied to the short-term memorability scores to account for varying memory retention durations. This was necessary because each video's successive presentation occurred after different time intervals (ranging from 45–100 intervening videos) introducing variability in recall probability. Since prior analysis of video memorability had shown that recall probability decays approximately linearly with increasing retention duration, this correction aimed to normalise scores by extrapolating memory performance to the maximal interval (100 videos). Notably, while linear decay provided the best fit for video data, image memorability has been better modeled using a log-linear decay function, as reported in the LaMem study (Khosla et al., 2015). The linear correction was not applied to long-term scores, where no consistent relationship between retention interval and raw scores was observed, and it was discontinued entirely after the 2020 edition.

In 2021, a new memorability score was introduced that employed an updated normalisation procedure grounded in the expected linear decay of recognition hit rates over short retention intervals. This revised approach allowed for the modelling of individual video-specific decay patterns by incorporating an adjustable decay-rate parameter, denoted as α, which accounts for subtle variations in how rapidly recognition performance declines across different videos. This video-specific decay rate is computed by assigning an α value to each video, obtained by minimizing the L_2 norm between the raw binary annotations and the predicted memorability score under a linear decay assumption, thereby yielding a decay parameter for each individual video. Full methodological details are provided in Newman et al. (2020). This normalisation procedure was subsequently adopted in the 2022 and 2023 editions of the task.

4 Techniques Used

In this section, we provide an overview of the techniques employed by participants in the MediaEval Predicting Video Memorability task across the 2018–2023 editions. These techniques span a wide range of approaches, including traditional machine learning, deep learning architectures, multimodal fusion strategies, and, more recently, brain-computer interface methods using EEG data.

To facilitate a structured and consistent discussion of these methods, we introduce a unified notation to refer to specific subtasks and configurations. Throughout the rest of this paper we will use the following notation for defining the memorability subtasks: *year.challenge.dataset.memorability_type* (see Figure 1). The *year* is defined as the year of the MediaEval competition, *challenge* is defined as one of the three main research directions, namely "P" for Prediction, "G" for Generalisation, and "E" for EEG, *dataset* signifies the dataset used in the experiments, while *memorability_type* symbolises the type of memorability scores used in the subtask, namely short- or long-term (abbreviated as ST and LT), raw scores or normalised (abbreviated as "norm") scores.

4.1 Analysis of the Overall Performance

We analyse general performance trends across existing approaches to the prediction tasks. Figure 7 presents a boxplot summarising the performance of submitted methods across different subtasks. Due to the limited number of submissions to the generalisation challenge type, our analysis is restricted to prediction-based subtasks. Results are grouped into five dataset and memorability type categories: VideoMem.ST, VideoMem.LT, TRECVid.ST, TRECVid.LT, and Memento.ST. As the prediction challenge was not part of the 2023 edition, this analysis spans the years 2018 to 2022. Across all analysed groups, we observe 13 negative outliers, suggesting only occasional poor-performing or misaligned runs.

We also observe notable trends in performance evolution across different editions of the task. Median performance values show consistent improvement across all dataset-memorability type combinations: VideoMem.ST (0.355 to 0.450), VideoMem.LT (0.168 to 0.203), TRECVid.ST (0.075 to 0.201 and 0.193), TRECVid.LT (0.043 to 0.060), and Memento.ST (0.602 to 0.604). These gains are substantial, with increases of 168% for TRECVid.ST, 26.7% for VideoMem.ST, and 20.8% for VideoMem.LT. Notably, even modest gains on Memento.ST are significant, as surpassing a Spearman rank correlation of 0.7 is particularly challenging, highlighting the strong performance of top submissions.

Top-performing systems also demonstrate consistent year-on-year improvement. Notable examples include: Video Mem.ST, with an increase from 0.497 Gupta and Motwani (2018) to 0.528 Azcona et al. (2019), TRECVid.ST from 0.137 Sweeney et al. (2021b) to 0.297 Gabriel Constantin and Ionescu (2022a), and Memento.ST from 0.658 Reboud et al. (2022) to 0.707 Agarla et al. (2023). Across datasets annotated for both short- and long-term memorability, systems generally perform better on short-term memorability.

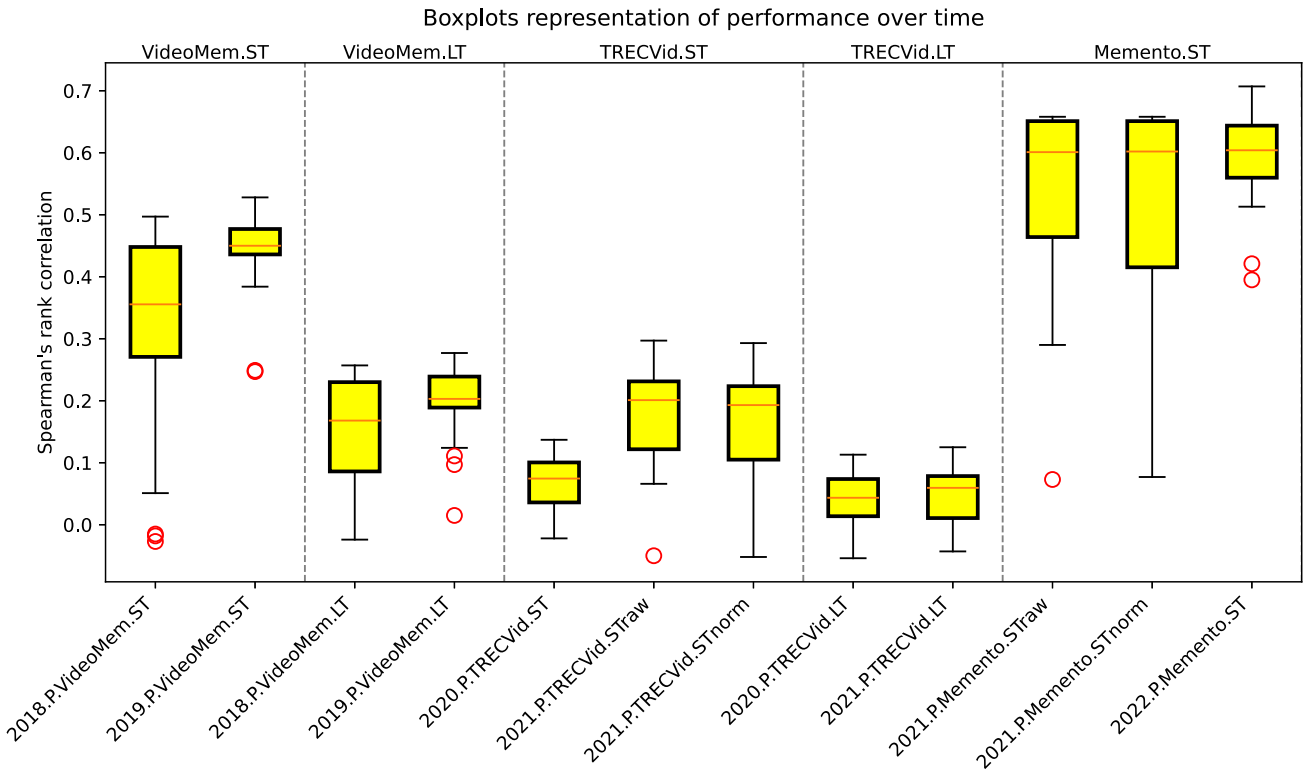


Fig. 7 Boxplot representation of the overall performance of methods across various versions of the dataset, for the prediction challenge. We use the following parameters in creating the boxplot representation: interquartile range $IQR = 50\%$, median values represented by the red lines, lower and upper adjacent values computed as $Q1-1.5\times IQR$ and $Q1+1.5\times IQR$, and outliers presented as red circles. *These results are for prediction-based subtasks only, results for the generalisation and EEG based subtasks will be presented separately.*

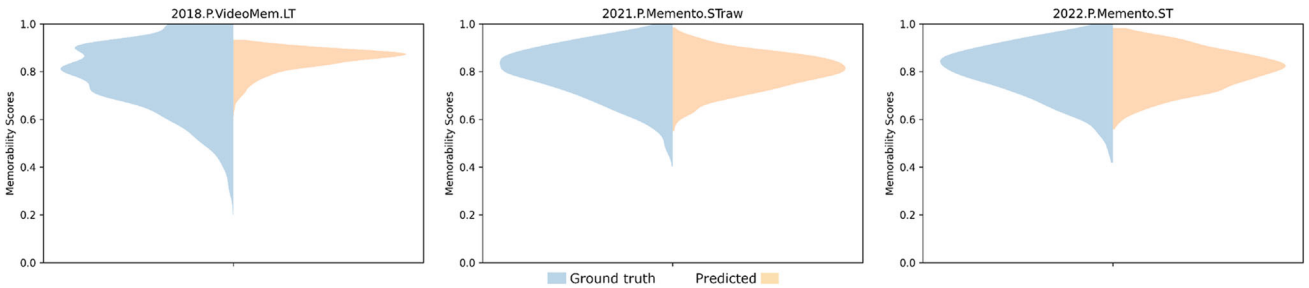


Fig. 8 A comparative analysis of the ground truth values and memorability prediction scores for the top performing systems in 2018. P.VideoMem.LT, 2021.P.Memento.STraw, 2021.P.Memento.STnorm. This figure presents a split violin plot comparing ground truth score distribution in blue with prediction score distribution in orange.

This consistent pattern reflects the greater difficulty in modelling long-term memory retention.

An analysis of submission behaviour reveals a consistent trend in the distribution of predicted scores. Figure 8 compares the distribution of ground truth memorability scores to that of the predicted scores from top participant submissions. Notably, participant predictions tend to compress the memorability range, producing a narrower score distribution than the ground truth. Although Figure 8 shows only three representative top-performing runs (2018.P.VideoMem.LT, 2021.P.Memento.STraw, and 2021.P.Memento.STnorm), similar trends were observed across the vast majority of runs and subtasks throughout all editions of the task.

4.2 Analysis of the Employed Features and Modalities

We now analyse the impact of the content descriptors and features employed by participants throughout the editions

of the Predicting Video Memorability task. Our goal is to identify the most commonly used feature types and assess how their prevalence correlates with system performance.

To support participants—particularly those early in their research career—we computed and provided the following set of pre-computed features:

- image-level features: FC7 activations from AlexNet (Krizhevsky et al., 2017), VGG (Simonyan and Zisserman, 2014), DenseNet121 (Huang et al., 2017), ResNet50 (He et al., 2016), EfficientNetB3 (Tan and Le, 2019); hand-crafted features including HOG (Dalal and Triggs, 2005), HSV and RGB histograms, and LBP (He and Wang, 1990);
- video-level features: C3D (Tran et al., 2015);
- audio features via the VGGish (Hershey et al., 2017) network for the datasets based on TRECVid;
- a set of text- and metadata-based features and information associated with each video, as provided by the creators of the original datasets (Cohendet et al., 2019; Awad et al., 2019; Newman et al., 2020).

A noteworthy trend across all editions is the widespread adoption of multimodal fusion strategies. Specifically, 48% of submitted runs employed early fusion techniques, while 24% utilised late fusion. Additionally, 9% of submissions integrated both early and late fusion within ensemble frameworks. These ensemble approaches were typically designed to enhance the output of single-modality systems and often yielded superior results. In fact, ensemble-based systems achieved top performance across nearly all short- and long-term prediction subtasks on the VideoMem and TRECVid datasets, with the Memento dataset being the only exception.

The best performing system on the short-term VideoMem data was developed by Azcona et al. (2019), for the 2019.P.VideoMem.ST subtask, with a result of 0.528. In this work, the authors employed a weighted ensemble strategy that fused the results of systems processing some of the visual pre-computed features, aesthetic and emotional-based features, and textual information. Their approach has also scored high on the 2019.P.VideoMem.LT subtask (0.27), but was not the top performing system. The top performing system in this case was presented by Reboud et al. (2019), using a weighted ensemble over visual and caption-based features.

For the short-term TRECVid subtasks, the top result was recorded in the 2021.P.TRECVid.STnorm subtask by the THAU-UPM team (Kleinlein et al., 2022), who achieved a score of 0.293 using a system that incorporated features from CLIP, ViT, and BEIT visual models. The same team also led the 2021.P.TRECVid.LT subtask, where they shifted focus to the textual modality, employing caption-based representations extracted via SBERT, GPT-2, and CLIP-text networks.

The top result in the Memento short-term subtask (2022.P.Memento.ST) was achieved by Agarla et al. (2023), who scored 0.707 using visual features derived from a ViT-style architecture.

Beyond the standard modalities, several participants experimented with alternative data modalities. A notable early example is the work of Smeaton Smeaton et al. (2018), who were the first to incorporate EEG signals into a memorability prediction system – a line of investigation that later gave rise to the dedicated EEG subtask. Saliency-based modelling also attracted interest, with both Smeaton (Smeaton et al., 2018) and Joshi et al. (2018) exploring visual attention mechanisms to inform memorability predictions.

Finally, we attempt to define and examine a set of modalities in order to determine whether any of them show consistent improvements in performance compared to others. We therefore explored the following categories: (i) visual image-level modalities, analyzing single-frame features; (ii) audio modalities; (iii) temporal modalities,encoding motion or temporal information in the visual domain; (iv) textual modalities, including provided and automatically generated captions; (v) conceptual modalities, capturing high-level attibutes such as aesthetics, emotional content, or perplexity; (vi) annotation-based modalities, involving models and methods that analyze annotations; and (vii) physiological modalities, using EEG-derived features. Although this analysis did not reveal a single modality or feature set that consistently outperforms others, it provides useful interpretability insights. The large variability of submitted systems within each modality and the prominence of early- and late-fusion approaches make it difficult to isolate the contibution of individual modalities. As we discuss further in Section 4.7, failure–analysis experiments indicate that certain visual and structural properties of videos (e.g., contrast, motion, object presence) correlate with prediction difficulty, offering initial evidence of recurring patterns linked to memorability. Addressing modality contributions more systematically remains an important direction for future work.

4.3 Analysis of the Classification Methods

In this section, we examine the techniques employed by participants. Our goal is to assess the capabilities of their methods and uncover general trends in their performance. In spite of the diversity of experimental approaches submitted over the years, discernible patterns have surfaced. While the categories presented below are not mutually exclusive, our intention is not to classify them but to ascertain the most salient approaches, explore their effectiveness, and gain insights into their performance and overarching tendencies.

We identify the following main approaches: (i) *Traditional Regressors*, (ii) *Neural Networks* (NN), (iii) *Deep Neural*

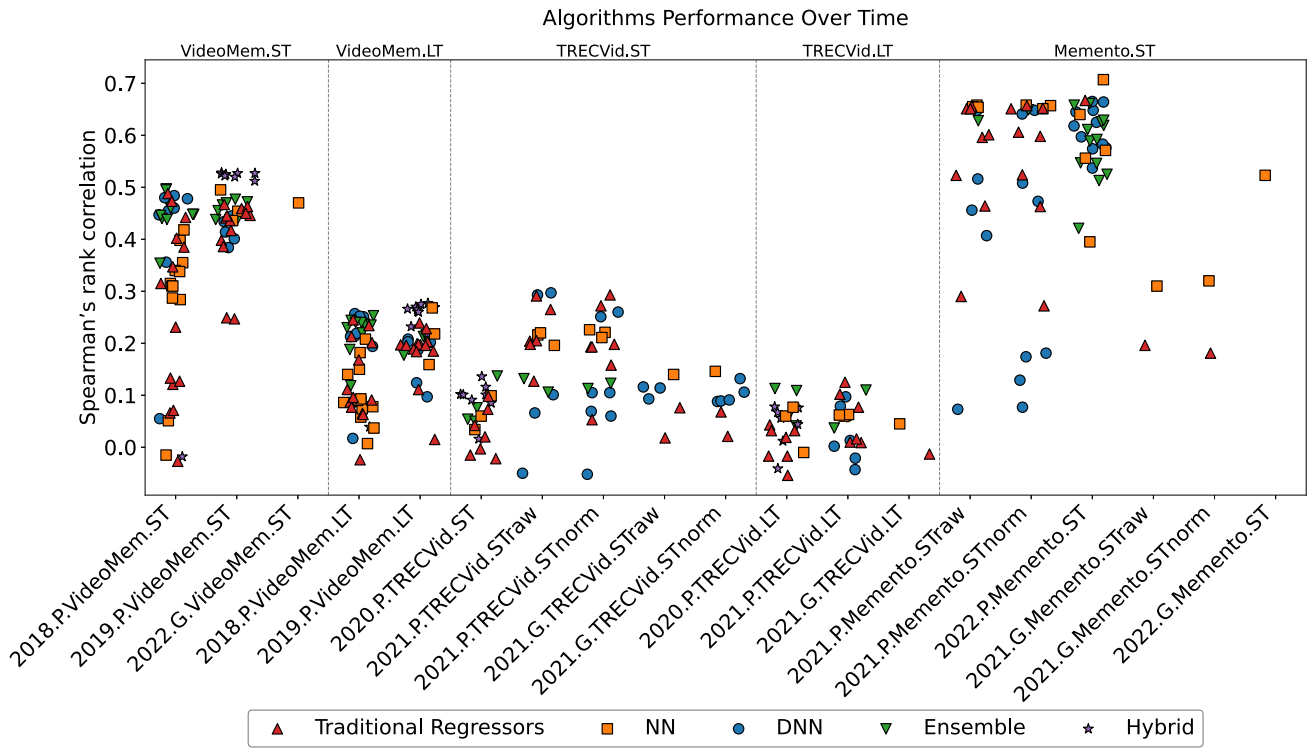


Fig. 9 Analysis of the employed methods for all the prediction and generalisation subtasks. We plot Spearman's rank correlation for all methods.

Networks (DNN), (iv) *Ensemble* i.e., combining the predictions of multiple individual models of the same category to produce a final prediction, and (v) *Hybrid* i.e., combining more than one approach of different categories. Please note that some of the categories might seem to include each other, e.g., deep networks and neural networks. In our analysis, "neural networks" refer to shallow networks with 3 or fewer hidden layers, while "deep networks" refer to end-to-end deep networks with more than three hidden layers. Results are presented in Figure 9.

Traditional Regressors. Traditional regressors were the favoured option across submissions, accounting for 31.8%. Notably, the top-ranking results on the 2021.P.TRECVid.STnorm and on the 2021.P.TRECVid.LT subtasks were achieved utilising this classification category. Both of these results were obtained by Kleinlein et al. (2022). The authors employed a Naïve Bayes regression model achieving a Spearman's rank correlation of 0.293 and 0.125, respectively. Traditional regressors were also strong candidates on other tasks. For example, Sweeney et al. (2023) employed a Bayesian Ridge Regressor model trained on stacked CLIP visual embeddings extracted from the first, middle, and last video frames, achieving a Spearman's rank correlation score of 0.667 on the 2022.P.Memento.ST subtask. Gupta and Motwani (2018) used an ElasticNet model that was trained using binary labels sampled from a binomial distribution, which were parameterised by the memorability scores. The motivation behind this approach stems from the fact that the memorability score indicates the percentage of subjects who can recall the contents of the video. As a result, the aforementioned technique achieved a Spearman's rank correlation score of 0.488 on the 2018.P.VideoMem.ST subtask.

Deep Neural Networks. DNNs were the second most used approach among the submitted classification methods, accounting for 22.6% of the total. They are used by the top runs on the 2018.P.VideoMem.LT, 2021.P.TRECVid.STraw and 2023.G.VideoMem.ST subtasks. Duy-Tue Tran-Van et al. (2018) used a recurrent neural network (RNN) with a single long short-term memory (LSTM) layer where each feature vector extracted from every video frame served as an input at a time step in the LSTM model. A dense layer was positioned at the top to take a 1024-dimensional output vector from the LSTM model to compute the video's memorability score. The approach achieved a Spearman's rank correlation score of 0.257 on the 2018.P.VideoMem.LT task. Constantin and Ionescu (Gabriel Constantin and Ionescu, 2022a) implemented a vision transformer architecture with a dedicated regression head that received filtered frames denoted as "Memorable Moments" selected based on data annotations to characterise a video from a memorability perspective. The approach achieved a Spearman's rank correlation score of 0.297 on the 2021.P.TRECVid.STraw subtask. The authors subsequently refined their approach in Gabriel Constantin and Ionescu (2023b) by augment-

ing the previously developed "Memorable Moments" to incorporate several video segments that corresponded to the instants most highly rated by annotators as memorable, in contrast to the previous approach of selecting only one segment per video. This allowed the authors to capture more representative video segments leading to the third-best run on the 2022.P.Memento.ST subtask, with a Spearman's rank correlation score of 0.665. The DNN approach presented by the SELAB-HCMUS team (Nguyen et al., 2024) used a series of fully connected layers for processing textual features extracted via CLIP, and visual features extracted via ResNet and EfficientNet. Each feature vector was processed by a set of 3 fully connected layers, followed by a final fully connected layer that joins all three branches. This approach scored the top result on the 2023.G.VideoMem.ST subtask, with a final result of 0.445. Additional noteworthy approaches comprise the work of Sweeney et al. (2022b), who used an xResNet50 model pretrained on ImageNet and fine-tuned on the Memento10k training data to obtain the second-best performance on both the 2021.G.TRECVid.STraw and 2021.G.TRECVid.STnorm subtasks with Spearman's rank correlation scores of 0.116 and 0.132 respectively.

Neural Networks. Shallow NN-based methods account for 18.3% of the submitted systems. They are used by two methods securing the first position on ten subtasks (Reboud et al., 2022) use textual features by concatenating video descriptions such as TF-IDF, GloVe embeddings, BERT token representations, and Sentence-BERT sentence representations. Additionally, visual features such as 2048-dimensional I3D features are extracted from the Mixed_5c layer trained on the Kinetics-400 dataset, and audio features obtained from the VGGish network that encode occurrence probabilities of the classes from the Google AudioSet Ontology. These multimodal feature vectors are passed as input to a feed-forward network that comprises a single hidden layer, where ReLU and dropout mechanisms are incorporated between the layers. The output for the regression result is generated using a simple sigmoid function. This approach yields the best results on various subtasks, including 2021.G.TRECVid.STraw, 2021.G.TRECVid.STnorm, 2021.G.TRECVid. LT, 2021. P.Memento.STraw, 2021.P.Memento. STnorm, 2021.G. Memento.STraw and 2021. G.Memento.STnorm, achieving Spearman's rank correlation scores of 0.14, 0.146, 0.045, 0.658, 0.658, 0.31 and 0.32, respectively. A similar approach is employed by (Agarla et al., 2023) which use a stack of three linear layers with ReLU in between and a sigmoid function for the regression result. The authors demonstrate that features extracted from a model trained with natural language supervision are effective for estimating video memorability obtaining the first place on the 2022.P.Memento.ST, 2022.G.Memento.ST, and 2022.G. VideoMem.ST subtasks with a Spearman's rank correlation score of 0.707, 0.523, and 0.47 respectively.

Ensemble. Ensemble methods account for 16.9% of all systems evaluated and have been central to several top-performing approaches, particularly in the 2018.P.Video Mem.ST, 2020.P.TRECVid.ST, and 2020.P.TRECVid.LT subtasks. In their 2018 submission, Gupta and Motwani (2018) employed a traditional ensemble of regression models, combined via a simple weighted averaging strategy. The optimal weights were selected through grid search over the development set. Their approach achieved a Spearman's rank correlation score of 0.497 on the 2018.P.VideoMem.ST subtask. Sweeney et al. (2021b) proposed a novel multi-stream ensemble framework centred on high-level conceptual audio features, using models like PANNs and xResNet34 for computing imageability, audio classification, and human causal uncertainty and arousal, as well as captions and visual features.

Hybrid. Hybrid approaches account for 10.3% of the analysed runs. They have been adopted by two leading approaches on the 2019.P.VideoMem.ST and 2019.P.Video Mem.LT subtasks. Azcona et al. (2019) used traditional regression models, including Support Vector Regression and Bayesian Ridge Regression, alongside deep neural networks such as VGG16, DenseNet121, ResNet50, and ResNet152. The models were fed with 8 frames from each source video, comprising the first frame and one frame from each of the seven seconds of the video. The authors employed an aggregation of model weights to construct the final ensemble model, where each weight represents the contribution of an individual model to the final prediction. To derive the optimal combination of weights, the authors employed a technique called "with replacement", which involves testing all possible combinations of weights by selecting values from a set of 20 bins, each containing 5% increments. Through exhaustive testing of all possible combinations of weights, the authors identified the combination that yielded the best results for the given dataset. This approach achieved the best results on the 2019.P.VideoMem.ST subtask with a Spearman's rank correlation score of 0.528. Meanwhile, Reboud et al. (2019) used a combination of traditional regression models, neural networks, and deep neural networks, which were separately trained on the visual and textual representations of videos, and subsequently fused using a weighted average method. This technique culminated in achieving the highest performance on the 2019.P.VideoMem.LT subtask, with a Spearman's rank correlation score of 0.27.

The prediction of video memorability is a challenging task that entails taking into account a broad range of factors that contribute to the memorability of a video. Videos are inherently multimodal, meaning that they contain information from multiple sensory modalities such as visual, auditory, and textual. Therefore, methods that are capable of han-

Table 3 Distribution of methods by dataset-memorability type pair. The percentage values in the table represent the proportion of models that used each method in each pair. The years in parentheses indicate the years when the datasets were used in the Predicting Video Memorability task.

	VideoMem.ST	VideoMem.LT	TRECVid.ST	TRECVid.LT	Memento.ST
Traditional Regressors	32.9%	32.9%	33.8%	33.3%	26.4%
Deep Neural Networks	17.1%	17.1%	28.2%	16.7%	33.3%
Neural Networks	19.5%	19.5%	15.5%	16.7%	19.4%
Ensemble	19.5%	17.1%	11.3%	14.3%	20.8%
Hybrid	11.0%	13.4%	11.3%	19.0%	0.0%
Years of activity	(2018,2019,2022,2023)	(2018,2019)	(2020,2021)	(2020,2021)	(2021,2022)

dling and integrating information from multiple modalities are likely to be more effective in predicting video memorability. Among the various types of methods that have been used for this task, neural networks, ensemble approaches, and hybrid systems have been found to be the most effective. These methods are able to learn and model complex relationships between different modalities and capture the underlying patterns and structures in the data. The superior performance of these methods can be attributed to their ability to handle the inherently multi-modal nature of videos and effectively integrate information from different modalities to make accurate predictions.

We further analyse in Table 3 the distribution of methods by dataset to determine the prevalence of different modelling approaches across the five dataset and memorability type pairs (VideoMem.ST, VideoMem.LT, TRECVid.ST, TRECVid.LT, and Memento.ST). The percentage of models that employed each method was calculated, and the results indicated varying trends across datasets. Traditional regressors were the most commonly used method, representing 32.9% of models in VideoMem.ST and VideoMem.LT, and 33.3% in TRECVid.LT. Deep neural networks were also prevalent, representing 33.3% of models in Memento.ST and 28.2% on TRECVid.ST. Ensemble methods were found to be less popular, representing only 11.3% of models in TRECVid.ST. A decreasing trend in the use of traditional regressors was observed in TRECVid.ST as these methods may have been superseded by newer, multimodal approaches, while the use of DNNs increased in TRECVid.LT and Memento.ST as these models have become increasingly powerful in recent years. Furthermore, neural networks, whether shallow or deep, constitute the most represented class of methods across all subtasks, datasets, and types of memorability. We can also observe an increase in the representation of the hybrid models in the TRECVid.LT. Overall, the results suggest that traditional regressors and DNNs remain popular modelling approaches, while the use of hybrid models is gaining popularity on certain datasets.

Another important aspect that must be discussed is the participants' approach to visual–temporal modelling. Two main directions emerged in this context. The first is a simple image-level strategy, in which participants analysed videos frame by frame and devised simple strategies such as averaging multiple frames or, in an even simpler variant, selecting a single representative frame to characterise the video. The second is a true temporal approach, where systems directly process temporal information using three-dimensional or sequence-based arquitectures. In this category, participants employed a variety of models, including LSTMs and recurrent network architectures (Duy-Tue Tran-Van et al., 2018), three-dimensional convolutional networks such as C3D (Le-Khac et al., 2021), I3D (Constantin et al., 2020), and, more recently, video vision transformer–based systems (Martín-Fernández et al., 2024). Table 4 presents a comparison between image- and video-based approaches.

Table 4 Result analysis summary for image-level vs video-level data processing throughout the subtasks, showing median (MED), maximum (MAX) and total number of runs (nmb) for each category.

	image			video		
Subtask	MED	MAX	nmb	MED	MAX	nmb
P.VideoMem.ST	*0.443*	0.522	54	0.355	**0.528**	19
P.VideoMem.LT	*0.201*	**0.277**	54	0.153	0.270	20
P.TRECVid.ST	0.101	**0.297**	36	*0.196*	0.226	11
P.TRECVid.LT	0.056	0.109	27	*0.061*	**0.110**	6
P.Memento.ST	0.603	0.661	30	*0.651*	**0.707**	17

First, it is easily noticeable that video-based approaches are considerably less represented among the submitted systems. However, when examining the results, their performance is broadly comparable to those of image–based methods. Higher maximum scores are achieved by image-level systems for P.VideoMem.LT and P.TRECVid.ST, while video-level processing achieves higher maximum scores for P.VideoMem.ST, P.TRECVid.LT, and P.Memento.ST. Median results are distributed in a similar fashion between the two approaches. Although, in the latest version of the task (P.Memento.ST), video-level approaches surpassed image-level approaches, overall variability prevents drawing a

definitive conclusion about the superiority of video–level modelling. We attribute this, at least in part, to the nature of the videos in the collection, which consist of short clips with limited temporal variation, for which a single frame or a small set of frames may already capture most of the relevant information.

4.4 Analysis of the Generalisation Subtasks

Throughout the editions of the task, we proposed and ran several generalisation subtasks, where users were asked to train their systems on one dataset, and test them on another. In this way, we wish to test the generalisation power of the systems while also making sure that they are subjected to a higher degree of novelty with regards to the testing data. We would like to clarify that, during the 2018–2023 period, the testing sets used across various datasets, editions and subtasks were never released to the public. Although certain datasets were no longer used in the prediction challenges, we retained their test sets and withheld their ground-truth labels in order to reuse them for generalisation challenges. Furthermore, researchers outside the MediaEval framework who wished to use these datasets were required to follow the same evaluation protocol as MediaEval participants: they did not have access to the test-set ground truth, but could submit their prediction files, which we evaluated and scored on their behalf.

A comparison between the results recorded by participants in the prediction subtasks vs. the generalisation subtasks is presented in Figure 10. While we did expect that generalisation systems would show lower performance compared with the direct prediction systems, we are pleasantly surprised to report that, in one case, the difference between prediction and generalisation approaches is not too high. This case is represented by the VideoMem.ST based approaches, where we find a decrease of only 11% between the top-performing prediction approaches and the top performing generalisation approaches. While the generalisation of any machine learning task is always an open research question, we can conclude that more experiments are needed for memorability generalisation, in order to draw a fair conclusion. Conversely, there are some major performance decreases in TRECVid.ST and Memento.ST, although in the latter case very few models have been submitted to the G.Memento.ST subtasks.

Table 5 presents an in-depth analysis of the generalisation subtasks, listing all subtasks, the testing datasets used, and the datasets involved in training for each case. It also provides a comparison between the performance of top runs on the generalisation and prediction challenges. Another noteworthy trend here is that approaches where data from TRECVid is included show a greater decrease in performance – 2021.G.TRECVid.ST and 2021.G.TRECVid.LT using TRECVid as the testing dataset, and 2021.G.Memento.

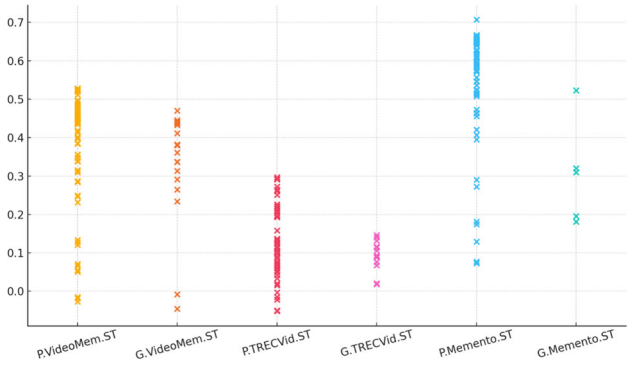


Fig. 10 Performance comparison between the prediction (P) and generalisation (G) challenges on the VideoMem.ST, TRECVid.ST, and Memento.ST datasets.

ST using TRECVid as the training dataset. All these approaches show a decrease in performance of over 50%, while the decrease for the other three generalisation challenges in under 30%. This may be an effect of the lower number of annotators on the TRECVid dataset, while also not excluding the possibility of a larger similarity between the visual data included in Memento and VideoMem.

4.5 Analysis of the EEG Subtasks

There were a total of three submissions for the EEG tasks: two in 2022 and one in 2023. While the EEG component was first introduced in 2021 as a pilot study, no participant submissions were received that year. The 2021 pilot focused on assessing the feasibility of using EEG signals to predict short-term memorability, and its findings informed the design of subsequent editions.

In the 2022 edition,Hamelink (2023) focused on predicting memorability using event-related potentials (ERPs) from the EEG data. The author focused on P300 components occurring up to 1000 ms after stimulus onset over the visual cortex and the right temporal lobe. Using linear mixed-effects models the authors investigated interactions between memorability and ERP amplitude. Kleinlein et al. (2023) investigated two different feature sets: summary statistics of each trial (mean, standard deviation, median, etc.), which were used to train random forest classifiers, and coherence maps computed across four canonical EEG frequency bands (delta, theta, alpha, beta) which were used to train a shallow neural network, achieving an AUC performance of 0.509.

The 2023 submission (Nguyen et al., 2024) extracted time and time-frequency domain features. Their best-performing method focused on accounting for individual differences using subject-specific one-hot encoding, achieving an AUC performance of 0.657. These submissions demonstrate the growing interest in leveraging EEG data for memorability prediction and highlight the methodological diversity emerg-

Table 5 Analysis of the generalisation subtasks throughout the years, showing individual subtasks, the datasets that were used for training and testing, the top performance of the testing dataset for the generalisation challenge in that subtask, and for the prediction challenge type, as well as the decrease in performance between the prediction and generalisation challenges.

testing set	subtask	training set	max P performance	max G performance	Decrease %
TRECVid	2021.G.TRECVid.ST	Memento	0.297	0.146	50.84%
	2021.G.TRECVid.LT	Memento	0.125	0.045	64%
Memento	2021.G.Memento.ST	TRECVid	0.658	0.32	51.36%
	2022.G.Memento.ST	VideoMem	0.707	0.523	26.02%
VideoMem	2022.G.VideoMem.ST	Memento	0.528	0.47	10.98%
	2023.G.VideoMem.ST	Memento	0.528	0.445	15.71%

ing in this subtask. As the field matures, future editions, such as the 2025 and 2026 editions, have further refined the experimental protocol and explore more robust, generalisable approaches to brain-based media analysis.

4.6 Future Trends

Although the methodological advances discussed in this section go beyond the official scope of the MediaEval Predicting Media Memorability task, recent developments in Artificial Intelligence, particularly the rise of Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) may substantially influence future research in memorability prediction.

Early explorations already point in this direction. Si et al. (2025) propose a fine-tuned LLaMA 13B model that uses as additional input a cognitive verbalization module describing the scene in detail, referring to scene clutter, object presence, aesthetic concepts, company logos, emotional content, and several other factors. This enriched semantic method achieves a performance of 0.5 on the 2021.P.TRECVid.ST dataset; however, in this case, performance is not computed on the official MediaEval testing set. More recently, Martín-Fernández et al. (2025) adapted a Qwen-VL-based model to memorability prediction by feeding three key images together with five captions as input for each video in the 2022.P.Memento.ST collection. Their proposed model surpasses all MediaEval submissions, achieving a score of 0.744, while using the official test set and methodology.

These advances reveal two complementary methodological directions for the integration of LVLMs into memorability prediction. The first leverages existing metadata attached to the video sample itself in order to augment the input to LVLMs, such as captions associated with the videos, video descriptions, subtitles, and OCR-extracted text. The second relies on automatically generated additional textual information, such as detailed video captions, object presence and object descriptions, as well as more subjective information like aesthetic and stylistic descriptions. While it remains an open question which type of information contributes most, current evidence suggests that combining both types of additional input data may be particularly beneficial, as it allows LVLMs to exploit both grounded and generated cues related to memorability

Furthermore, while newer iterations of the MediaEval memorability task are not analyzed in this work due to the significant shift in the task's scope, we expect LLM- and LVLM-based systems to become increasingly prominent in forthcoming editions of the MediaEval task. Future iterations will include longer and more semantically rich content, such as movie excerpts and commercial videos. Such material often contains domain-specific language, explicit narrative structure, and product- or brand-related messaging, making textual modalities–especially transcripts and high-quality captions–particularly informative. As a result, models capable of integrating visual, auditory, and linguistic representations in a unified multimodal architecture are likely to play a central role in advancing the state of the art.

4.7 Failure Analysis Experiments

Following the 2022 edition of the MediaEval Predicting Video Memorability task, we conducted an in-depth failure analysis using the complete set of submissions for the 2022.P.Memento.ST subtask, the last prediction challenge type subtask in our collection, and a set of features and video characteristics. As detailed in our previous work (Gabriel Constantin et al., 2023), we introduced a metric designed to quantify how easy it is for the models submitted by participants to accurately measure the memorability of a video. This metric takes into account the ground truth memorability scores associated with each video $G = \{g_1, g_2, ..., g_N\}$, and a set of predictions composed of all the predictions from the participants' submissions for each video i, $P_i = \{p_1^i, p_2^i, ..., p_M^i\}$. Thus, a "difficulty distance" can be measured for each video, as the sum of the absolute differences between the predicted values and the ground truth for each video: $D_i = \sum \left| p_j^i - g_i \right|$, which can then be normalised in order to bring values in the [0, 1] interval.

The following step involves grouping the videos into four equal quartiles, as well as into four threshold-based intervals, with the threshold growing by 0.25 for each group of videos. Finally, we extracted a series of visual, object-based, and annotator-based features for each video, while looking at trends in the groups of videos. For a comprehensive list of features, we direct the reader to the paper Gabriel Constantin et al. (2023); here, we report the main findings most relevant to interpretability. Videos that were consistently difficult for systems to predict exhibit recurring properties: (i) higher contrast and sharpness; (ii) lower dynamism; (iii) lower brightness and colorfulness; (iv) higher saturation; (v) fewer objects in them and the objects occupy a smaller portion of the frames; (vi) ground-truth memorability scores in the mid-range, which appear inherently more ambiguous for models.

These observations offer initial evidence of which visual and structural attributes correlate with prediction difficulty–and, by contrast, which factors tend to facilitate more reliable memorability estimation. We believe that such analyses provide valuable interpretability insights, helping researchers identify which types of videos are under-represented or more challenging, and guiding more informed dataset construction, augmentation strategies, and model design in future work.

5 Super System Design

Given the success of ensemble systems shown during early editions of the Predicting Media Memorability task, we carried out further experiments which used the runs submitted by our participants while applying different late fusion approaches to their predictions. This was done in an attempt to achieve better results than individual systems (which, from this point on, will be referred to as *inducers*). One of the main drawbacks of systems such as these is represented by the significant demand for computational resources, i.e., in order to replicate this experiment, one would have to run all the inducer systems, extract their predictions, and then run the ensemble over those results. However, we carry out these experiments in the hope that it will give us a general idea with regards to a maximal possible soft-plateau of performance for memorability prediction systems on the datasets that were used throughout the editions of the Predicting Video Memorability task.

Some of the super systems presented here require an additional training step (the regressor systems, as well as weighted approaches that used biased sampling). In those cases, we created a five-fold cross-validation experiment, splitting the available inducer data between a training and a testing set, and averaging the performances recorded during the five experiments into a single value.

We used an extensive set of 14 methods of performing ensembling, that used the predictions of the inducers submitted by participants as inputs or features for describing a video sample. We start with simple, statistical approaches to ensembles, using the weighted mean, mean, median, minimum and maximum over the inducer predictions. We then followed with some more complex approaches, including linear, ridge, Huber, and Lasso Regressors, as well as boosting strategies like XGBoost, gradient boosting, SVR, random forest, and shallow neural networks. We did not filter out any of the submitted runs, as we wanted to determine if systems are able to adapt to inconsistencies present in the input space.

Results of these experiments are summarised in Table 6 where we report only the three best performing ensembles. It is interesting to conclude that, for all dataset and memorability type pairs, i.e., VideoMem.ST, VideoMem.LT, TRECVid.ST, TRECVid.LT, and Memento.ST, at least one of the ensemble systems developed for the Super System yielded results that surpass those of the individual inducers. The highest increase recorded across all datasets is recorded for TRECVid.LT, where an 18.4% increase is achieved from the top performing MediaEval system (with a score of 0.125) to the Weighted Mean (Biased Sampling) ensemble approach (with a score of 0.148). Finally, for this set of experiments, it would seem that weighted mean (biased sampling) is generally among the top ensemble performers.

6 Lessons Learned and Open Questions

Reflecting on the six editions of the MediaEval Predicting Video Memorability task, several key insights have emerged that inform both the design of future benchmarks and the broader understanding of computational memorability.

One of the most consistent observations across datasets with both short- and long-term annotations (namely the VideoMem and TRECVid datasets) is the superior performance of systems on short-term vs. long-term memorability tasks. For instance, the top performing systems achieved a Spearman's rank correlation of VideoMem 0.528 for VideoMem short-term, compared to 0.277 for long-term; and 0.297 vs. 0.125 for TRECVid. This discrepancy is likely due to the inherent complexity of long-term memory. While short-term memory reflects recalling the videos a viewer sees within the same annotation session, long-term memorability requires viewers to remember videos after a longer period of time (24–72 hours in our case), introducing additional cognitive and contextual variables that complicate prediction. Moreover, long-term annotations typically involve fewer annotators when compared with short-term, as they require participants to return for a second session. This makes data collection more challenging and potentially less robust. For example, the average number of annotators

Table 6 Summary of the Super System approaches. For each variation of the dataset and memorability type, we indicate in *italic* the best-performing inducer submitted by participants to the MediaEval tasks, as well as the three best performing ensemble methods. The best overall system is indicated in **bold**. Also, the asterisk indicates systems that did not need a training step. For fusion systems which were trained, we present the uncertainty estimates.

Dataset	System	Performance
VideoMem.ST	*top-MediaEval*	*0.528*
	Ensemble - SVR	**0.572 ± 0.023**
	Ensemble - Weighted Mean (Biased Sampling)	0.536 ± 0.018
	Ensemble - ElasticNet	0.531 ± 0.027
VideoMem.LT	*top-MediaEval*	*0.277*
	Ensemble - Ridge Regressor	**0.303 ± 0.019**
	Ensemble* - Min	0.296
	Ensemble - Weighted Mean (Biased Sampling)	0.294 ± 0.012
TRECVid.ST	*top-MediaEval*	*0.297*
	Ensemble - Weighted Mean (Biased Sampling)	**0.314 ± 0.021**
	Ensemble* - Weighted Mean	0.290
	Ensemble* - Median	0.277
TRECVid.LT	*top-MediaEval*	*0.125*
	Ensemble - Weighted Mean (Biased Sampling)	**0.148 ± 0.014**
	Ensemble* - Weighted Mean	0.118
	Ensemble* - Median	0.078
Memento.ST	*top-MediaEval*	*0.707*
	Ensemble - ElasticNet	**0.762 ± 0.019**
	Ensemble - Weighted Mean (Biased Sampling)	0.737 ± 0.026
	Ensemble - Lasso Regression	0.732 ± 0.027

for VideoMem short-term was 38 and only 13 for long-term; similarly, TRECVid had 22 annotators for short-term and just 7 for long-term. These differences likely contribute to the observed performance gap between the two tasks.

Another important observation relates to the number of annotators involved in the data collection process. Datasets with a higher number of annotators tend to yield better prediction performance. Looking at the results across different tasks, we observed that the top performing runs, with results going up to 0.707, are achieved for Memento.ST, a dataset benefiting from 90 annotators on average, significantly higher than either VideoMem or TRECVid. In contrast, the lowest results when comparing the three short-term datasets are achieved on TRECVid, the dataset with the lowest number of annotators. This suggests that a higher number of annotators helps in reducing the degree of subjectivity associated with the memorability judgments, and therefore helps improving the reliability of the ground truth and the accuracy of predictive models.

Another important lesson concerns the modelling strategies adopted by participants. Systems that employed fusion models (whether early, late, or hybrid) outperformed single-modality approaches. This trend reflects the inherently multimodal nature of video content, which combines visual, auditory, and textual information. Our experiments with an ensemble-based "Super System" further confirmed that combining predictions from multiple models can lead to improved performance. In every dataset examined, at least one ensemble method outperformed the best individual model submitted by participants. This suggests that different models capture complementary aspects of memorability, and that ensemble learning remains a promising direction for future research. A deeper analysis of the diversity among individual models (or "inducers") could reveal which features or modalities contribute most to these performance gains. Also, as discussed in Section 4.6, inherently multimodal approaches such as LVLMs show particularly promising performance for memorability prediction. Although most of these results have been reported outside the official MediaEval evaluations, several works have used the same data splits as the benchmark, and surpassed the best results recorded during the competition. This suggests that richer multimodal representations may help overcome some of the limitations observed in traditional approaches, while also indicating that the current benchmark conditions do not yet fully capture the capabilities of emerging models.

Generalisation is a challenging aspect of memorability prediction. While systems tested on VideoMem showed only a modest drop in performance when evaluated on unseen data (around 11%), the same was not true for the TRECVid and Memento datasets. While for Memento the case can be made that not many systems were submitted to the generalisation task, the same cannot be said about TRECVid, a dataset showing major decreases in top performance. Interestingly, the same phenomenon occurs when TRECVid is used as the training dataset, while Memento is used as the testing dataset.

The sharp decline in generalisation performance associated with TRECVid may be attributed to its lower annotation density and greater variability in content, which complicates model transferability. Additionally, the structure of the generalisation task – where models were trained on Memento and tested on TRECVid – may have exposed limitations in cross-domain learning. These findings highlight the need for more robust and diverse training data, as well as improved methods for domain adaptation.

Participation trends also offer valuable insights. Many of the generalisation runs for VideoMem were submitted during the 2023 edition, when only the generalisation and EEG subtasks were ongoing. The absence of the prediction task in that edition led to a noticeable increase in generalisation submissions, suggesting that participants tend to prioritise the main or most accessible task. This observation could inform future benchmark design by encouraging broader engagement across subtasks through balanced task emphasis or targeted incentives.

The EEG subtask, introduced as a pilot in 2021 and expanded in subsequent years, has shown encouraging signs of progress. Although participation remains limited, the submissions received in 2022 and 2023 demonstrate an interest in neurophysiological approaches to memorability prediction. The use of event-related potentials, coherence maps, and time-frequency features suggests a rich methodological landscape, and future editions of the task may benefit from deeper exploration of cross-subject generalisation and interpretability in EEG-based models.

As discussed above, our analysis did not reveal any modality or feature type that consistently outperforms others across editions or implementations. This reinforces the importance of analyzing videos in their entirety from a multimodal perspective. It remains to be seen whether this observation will hold in future iterations of the task as the nature of the data evolves. The 2018–2023 editions of the benchmark relied primarly on short, social media–type content, whereas videos in the upcoming editions will incorporate movie excerpts and commercial content, introducing richer narrative structure, domain-specific language, and more complex audio-visual dynamics. Such changes may shift the relative importance of different modalities. For instance, commercial videos often contain product-focused spoken or textual information, suggesting that transcripts or language-based features may become more predictive than purely visual or audio descriptors. While this remains a hypothesis and has not yet been confirmed through systematic experimentation, it illustrates how modality relevance may be highly dependent on content type.

In summary, the lessons learned from six years of benchmarking reveal both the progress made and the challenges that remain. They underscore the importance of robust annotation protocols, multimodal modelling, and thoughtful task design in advancing the field of computational memorability. Our main findings and the open questions that remain to be explored with regard to model and feature design can be summarised as follows:

- Multimodal and fusion-based approaches consistently achieve better results, reflecting the intrinsically multimodal nature of memorability and the heterogeneous cues present in short social media videos.
- No single modality, feature, or modelling approach emerges as universally superior across datasets or editions; to consistently outperform the others.
- Recent advances in large vision–language models show clear promise for memorability prediction, particularly when enriched with textual or semantic information, although their full potential within the benchmark setting remains to be systematically explored.
- The number of annotations per video is crucial for producing reliable ground truth and for enabling systems to achieve stable performance, with higher annotation density correlating with more predictable memorability estimation.
- For short-duration videos where scene content varies minimally, both image- and video-level analysis remain relevat;

In the upcoming editions we will continue to monitor these and other aspects of system design closely. As the nature of the videos will change significantly, as we pivot toward movie excerpts and commercial content, some of the observations drawn from the 2018–2023 cycles may evolve, and certain open questions may find clearer answers. At the same time, this transition may introduce new challenges for model generalizability, as systems trained on one content domain may not transfer reliably to others without additional adaptation. Addressing these domain-specific differences will likely require reconsidering both feature design and multimodal integration strategies for conceptually distinct categories of data.

7 Conclusions

This paper presents a comprehensive review of the MediaEval Predicting Video Memorability Task, covering its evolution from 2018 to 2023. Through an analysis of datasets, annotation protocols, modelling techniques, and participant submissions, we have traced the development of computational memorability as a research domain. The benchmark has provided a reproducible and collaborative framework for evaluating memorability prediction systems, fostering innovation and methodological diversity.

While significant progress has been made, particularly in short-term memorability prediction and multimodal modelling, several challenges remain. Long-term memorability, generalisation across datasets, and the integration of EEG data continue to pose open questions. The lessons learned from this benchmark offer valuable guidance for future research, emphasizing the importance of annotation scale, model fusion, and cross-domain robustness. As the field continues to evolve, we anticipate that future editions of the task, such as the 2025 and 2026 iterations, will further expand the scope of memorability research, incorporating longer-form content and more sophisticated modelling paradigms. By building on the insights gained from past editions, researchers can move closer to a comprehensive understanding of what makes media memorable.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Financial support was provided by the University of Essex Faculty of Science and Health Research Innovation and Support Fund and by Science Foundation Ireland under Grant Number SFI/12/RC/2289_P2, cofunded by the European Regional Development Fund. The work was partly funded through the YERUN Research Mobility Awards 2021. Financial support was provided under project grants PN-IV-P6-6.3-SOL-2024-3-0320 and PN-IV-P6-6.3-SOL-2024-0060, funded by UEFISCDI, Solutions Axis. This work was partly supported by the project GRESEL-UNED PID2023-151280OB-C22 funded by MICIU/AEI/ AEI 501100011033. This work was partly supported by the Spanish Ministry of Science, Innovation and Universities (project ANNOTATE (PID2024-156022OB-C31)) funded by MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+). The research of Iván Martín-Fernández was supported by the Universidad Politécnica de Madrid (Programa Propio I+D+i). This work was funded by the Spanish Ministry of Science and Innovation through the project TRUSTBOOST (PID2023-150584OB-C21), funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR". This work was partly supported by TüBİTAK under the 2209-A (Project No: 1919B012468481).

Data Availability All the data is available upon request to the authors of this paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Agarla, M., Celona, L., & Schettini, R. (2023). Predicting video memorability using a model pretrained with natural language supervision. In: *Proc. of the MediaEval 2022 Workshop*, vol.3583 of *CEUR Workshop Proc.*, Bergen, Norway, 2023.

Awad, G., Butt, A.A., Curtis, K., Lee, Y., Fiscus, J., Godil, A., Delgado, A., Zhang, J., Godard, E., Diduch, L., et al. (2019). TRECVID 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. In: *Proceedings of TRECVID 2019*. USA.

Azcona, D., Moreu, E., Hu, F., Ward, T. E., & Smeaton, A. F. (2019). Predicting media memorability using ensemble models. In: *Proc. of the MediaEval 2019 Workshop*, vol.2670 of *CEUR Workshop Proc.*, 27-30.

Cohendet, R., Demarty, C.-H., Duong, N. Q. K., Sjöberg, M., Ionescu, B., & Do, T.-T. (2018). MediaEval 2018: Predicting media memorability task. In: *Proc. of the MediaEval 2018 Workshop*, vol.2283 of *CEUR Workshop Proc.*, Sophia Antipolis, France.

Cohendet, R., Demarty, C. H., Duong, N. Q., & Engilberge, M. (2019). VideoMem: Constructing, analyzing, predicting short-term and long-term video memorability. *Proceedings of the IEEE International Conference on Computer Vision, 55*, 2531–2540.

Constantin, M. G., & Ionescu, B. (2022). Using vision transformers and memorable moments for the prediction of video memorability. In: *Proc. of the MediaEval 2021 Workshop*, vol.3181 of *CEUR Workshop Proc.*, Online.

Constantin, M. G., & Ionescu, B. (2023). Aimultimedialab at mediaeval 2022: Predicting media memorability using video vision transformers and augmented memorable moments. In: *Proc. of the MediaEval 2022 Workshop*, vol.3583 of *CEUR Workshop Proc.*, Bergen, Norway.

Constantin, M. G., Demarty, C.-H., Fosco, C., García Seco de Herrera, A., Halder, S., Healy, G., Ionescu, B., Matran-Fernandez, A., Savran Kiziltepe, R., Smeaton, A.F., & Sweeney, L. (2024). Overview of the MediaEval 2023 predicting video memorability task. In: *Proc. of the MediaEval 2023 Workshop*, vol.3658 of *CEUR Workshop Proc.*, Amsterdam, The Netherlands.

Constantin, M. G., Dogariu, M., Jitaru, A. C., & Ionescu, B. (2023). Assessing the difficulty of predicting media memorability. In: *Proceedings of the 20th International Conference on Content-based Multimedia Indexing*, pages 188–192.

Constantin, M. G., Ionescu, B., Demarty, C. H., Duong, N. Q., Alameda-Pineda, X., & Sjöberg, M. (2020) The predicting media memorability task at MediaEval 2019. In: *Proc. of the MediaEval 2019 Workshop*, vol.2670 of *CEUR Workshop Proc.*, Sophia Antipolis, France, Sep.

Constantin, M. G., Kang, C., Dinu, G., Dufaux, F., Valenzise, G., & Ionescu, B (2020). Using aesthetics and action recognition-based networks for the prediction of media memorability. In: *Proc. of the MediaEval 2019 Workshop*, vol.2670 of *CEUR Workshop Proc.*, Sophia Antipolis, France, Sep. 2020.

Dalal, N., & Triggs, B. (2005) Histograms of oriented gradients for human detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

De Herrera, A. G. S., Kiziltepe, R. S., Chamberlain, J., Constantin, M. G., Demarty, C. H., Doctor, F., Ionescu, B., & Smeaton, A. F. (2021) Overview of MediaEval 2020 predicting media memorability task: What makes a video memorable? In: *Proc. of the MediaEval 2020 Workshop*, vol.2882 of *CEUR Workshop Proc.*, Online, 2021.

Gupta, R., & Motwani, K. (2018). Linear models for video memorability prediction using visual and semantic features. In: *Proc. of the MediaEval 2018 Workshop*, vol.2283 of *CEUR Workshop Proc.*, Sophia Antipolis, France.

Gygli, M., Grabner, H., Riemenschneider, H., Nater, F., & Van Gool, L. (2013). The interestingness of images. In: *Proceedings of the IEEE international conference on computer vision*, pages 1633–1640.

Hamelink, R. (2023) Visual recognition of the memento10k dataset: A p300 component erp study. In: *Proc. of the MediaEval 2022 Workshop*, vol.3583 of *CEUR Workshop Proc.*, Bergen, Norway.

He, K., Zhang, X., Ren, S., & Sun, J. (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B. et al. (2017). Cnn architectures for large-scale audio classification. In: *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

He, D.-C., & Wang, L. (1990). Texture unit, texture spectrum, and texture analysis. *IEEE Transactions on Geoscience and Remote Sensing, 28*(4), 509–512.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017) Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitionn*, pages 4700–4708.

Isola, P., Parikh, D., Torralba, A., & Oliva, A. (2011) Understanding the intrinsic memorability of images. *Advances in neural information processing systems*, 24.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011) What makes an image memorable? In: *CVPR 2011*, pages 145–152. IEEE.

Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2013). What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence, 36*(7), 1469–1482.

Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., & Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. 8:e47596.

Joshi, T., Sivaprasad, S., Bhat, S., & Pedanekar, N. (2018) Multimodal approach to predicting media memorability. In: *Proc. of the MediaEval 2018 Workshop*, vol.2283 of *CEUR Workshop Proc.*, Sophia Antipolis, France.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015) Understanding and predicting image memorability at a large scale. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Kiziltepe, R. S., Constantin, M. G., Demarty, C. H., Healy, G., Fosco, C., de Herrera, A. G. S., Ionescu, B., Matran-Fernandez, A., Smeaton, A. F.,& Sweeney, L. (2022) Overview of the MediaEval 2021 predicting media memorability task. In: *Proc. of the MediaEval 2021 Workshop*, vol.3181 of *CEUR Workshop Proc.*, Online.

Kleinlein, R. (2023). Enrique Rodríguez-Sebastián, and Fernando Fernández-Martínez. Understanding media memorability from event-related potential records and visual semantics. In: *Proc. of the MediaEval 2022 Workshop*, vol.3583 of *CEUR Workshop Proc.*, Bergen, Norway.

Kleinlein, R., Luna-Jiménez, C., & Fernández-Martínez, F. (2022). THAU-UPM at MediaEval 2021: From video semantics to memorability using pretrained transformers. In: *Proc. of the MediaEval 2021 Workshop*, vol.3181 of *CEUR Workshop Proc.*, Online.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects.,*139*(3), 558–578.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: The role of categories in visual long-term memory. 21(11):1551–1556.

Krizhevsky, A., Sutskever,I., Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Le-Khac, P. H., Rai, A.K., Healy, G., Smeaton, A. F., O'Connor, N.E. (2021). Investigating memorability of dynamic media. In: *Proc. of the MediaEval 2020 Workshop*, vol.2882 of *CEUR Workshop Proc.*, Online.

Lin, Q., Yousif, S.R., Chun, M.M.,& Scholl, B.J. (2021). Visual memorability in the absence of semantic content. 212:104714.

Martín-Fernández, I., Romero, S.E., Bellver-Soler, J., Gil-Martín, M., & Martínez, F.F. (2024). Exploring video transformers and automatic segment selection for memorability prediction. In: *Proc. of the MediaEval 2023 Workshop*, vol.3658 of *CEUR Workshop Proc.*, Amsterdam, The Netherlands.

Martín-Fernández, I., Esteban-Romero, S., Fernández-Martínez, F., & Gil-Martín, M. (2025). Parameter-efficient adaptation of large vision-language models for video memorability prediction. *Sensors (Basel, Switzerland), 25*(6), 1661.

Newman, A., Fosco, C., Casser, V., Lee, A., McNamara, B., & Oliva, A. (2020). Multimodal memorability: Modeling effects of semantics and decay on video memorability. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 223–240, Cham.

Nguyen, M.-Q., Trinh, M.-H., Bui, H.-G., Vo, K.-T., Tran, M.-T., Tran, T.-P., & Nguyen, H.-D. (2024). SELAB-HCMUS at MediaEval 2023: A cross-domain and subject-centric approach towards the memorability prediction task. In: *Proc. of the MediaEval 2023 Workshop*, vol.3658 of *CEUR Workshop Proc.*, Amsterdam, The Netherlands.

Reboud, A., Harrando, I., Laaksonen, J., Francis, D., Troncy, R., & Laria Mantecón, H. (2019). Combining textual and visual modeling for predicting media memorability. In: *Proc. of the MediaEval 2019 Workshop*, vol.2670 of *CEUR Workshop Proc.*, Sophia Antipolis, France.

Reboud, A., Harrando, I., Laaksonen, J.,& Troncy, R. (2022). Exploring multimodality, perplexity and explainability for memorability prediction. In: *Proc. of the MediaEval 2021 Workshop*, vol.3181 of *CEUR Workshop Proc.*, Online.

Roger, R. N. (1967). Shepard. *Recognition memory for words, sentences, and pictures., 6*(1), 156–163.

Si, H., Singh, S., Singla, Y. K., Bhattacharyya, A., Baths, V., Chen, C., Shah, R. R., & Krishnamurthy, B. (2025). Long-term ad memorability: Understanding & generating memorable ads. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5707–5718. IEEE

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Smeaton, A.F., Corrigan, O., Dockree, P., Gurrin, C., Healy, G., Hu, F., McGuinness, K., Mohedano, E., & Ward, T.E. (2018). Dublin's participation in the predicting media memorability task at MediaEval 2018. In: *Proc. of the MediaEval 2018 Workshop*, vol.2283 of *CEUR Workshop Proc.*, Sophia Antipolis, France.

Standing, L. (1973). Learning 10000 pictures. 25(2):207–222.

Sweeney, L., Constantin, M.G., Demarty, C.-H., Fosco, C., Seco de Herrera, A.G., Healy, G., Halder, S., Ionescu, B., Matran-Fernandez, A., Smeaton, A.F., & Sultana, M. (2023). Overview of the MediaEval 2022 predicting video memorability task. In: *Proc. of the MediaEval 2022 Workshop*, vol.3583 of *CEUR Workshop Proc.*, Bergen, Norway.

Sweeney, L., Healy, G., & Smeaton, A.F. (2021). Leveraging audio gestalt to predict media memorability. In: *Proc. of the MediaEval 2020 Workshop*, vol.2882 of *CEUR Workshop Proc.*, Online.

Sweeney, L., Healy, G., & Smeaton, A.F. (2022). Predicting media memorability: Comparing visual, textual, and auditory features. In: *Proc. of the MediaEval 2021 Workshop*, vol.3181 of *CEUR Workshop Proc.*, Online.

Sweeney, L., Healy, G., & Smeaton, A.F. (2023). Diffusing surrogate dreams of video scenes to predict video memorability. In: *Proc. of the MediaEval 2022 Workshop*, vol.3583 of *CEUR Workshop Proc.*, Bergen, Norway.

Sweeney, L., Matran-Fernandez, A., Halder, S., Seco de Herrera, A.G., Smeaton, A.F., & Healy, G. (2022). Overview of the EEG pilot subtask at MediaEval 2021: predicting media memorability. In: *Proc.*

of the MediaEval 2021 Workshop, vol.3181 of *CEUR Workshop Proc.*, Online.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*, pages 6105–6114. PMLR.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp 4489–4497.

Tran-Van, D.-T., Tran, L.-V., Tran, M.-T. (2018). Predicting media memorability using deep features and recurrent network. In: *Proc. of the MediaEval 2018 Workshop*, vol.2283 of *CEUR Workshop Proc.*, Sophia Antipolis, France.

Xie, W., Bainbridge, W.A., Inati, S.K., Baker, C.I., & Zaghloul, K.A. (). Memorability of words in arbitrary verbal associations modulates memory retrieval in the anterior temporal lobe. 4(9):937–948 (2020).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.