Multimedia Tools and Applications (2026) 85:30
https://doi.org/10.1007/s11042-026-21260-3

A comprehensive study on contrastive pre-training and fine tuning of vision and text transformers for video memorability prediction

Iván Martín-Fernández[1] · Sergio Esteban-Romero[1] · Manuel Gil-Martín[1] · Fernando Fernández-Martínez[1]

Received: 7 November 2024 / Revised: 27 June 2025 / Accepted: 10 October 2025

Abstract
Video memorability prediction has emerged as a key challenge for improving information retrieval, content design, and user engagement. Prior work has shown that semantic cues play a crucial role in determining memorability, with recent studies leveraging Contrastive Language-Image Pre-training (CLIP) encoders to incorporate semantic information. However, the specific improvements attributable to CLIP models remain unclear, as few studies systematically compare their performance against equivalent unimodal encoders or explore fine-tuning strategies. This work addresses that gap through a comprehensive, controlled evaluation of CLIP-based and unimodal encoders for video memorability prediction. We propose FCLIP, a domain-adapted extension of CLIP that undergoes additional contrastive pre-training on memorability-specific image-text pairs. Our experiments assess both feature extraction and supervised fine-tuning, ensuring fair comparisons across models with matched architecture and parameter count. Results show that FCLIP image encoders achieve a Spearman Rank Correlation Coefficient (SRCC) of 0.672 on the Memento10k dataset, significantly outperforming unimodal Vision Transformers. FCLIP text encoders similarly outperform unimodal baselines, reaching an SRCC of 0.632. These findings demonstrate that contrastive learning and domain adaptation substantially improve memorability prediction, highlighting the importance of semantic and multimodal pre-training in developing advanced content analysis systems.

Keywords Video memorability prediction · Contrastive language image pre-training (CLIP) · Multimodal content analysis · Semantic knowledge integration

1 Introduction

In contemporary society, multimedia content has become an integral component of human communication. Consequently, identifying content that is genuinely relevant to a particular user or community has become a task of increasing relevance and complexity. The

Extended author information available on the last page of the article

Published online: 27 January 2026

development of automated systems capable of modeling the memorability of an audiovisual stimulus, or its inherent capacity to be remembered, offers a solution to the challenge of identifying the relevant information. Research in this domain has the potential to enhance comprehension of the human perception of multimedia content. Furthermore, it can directly impact the industry, leading to the development of more effective advertising campaigns, the creation of more impactful educational content, and an increase in the viewer's ability to retain the information presented. Considering this, the objective of this paper is to contribute to the advancement of research in the field of predicting the memorability of multimedia content by the effect of semantic learning techniques on the performance of these models.

Memorability, within the realm of human perception, a field at the intersection of philosophy, psychology, and neuroscience, examines how individuals process and remember environmental stimuli [1]. This process is multimodal, engaging multiple senses to interpret complex signals, and is influenced by cognitive functions such as attention, learning, and expectancy. Memory plays a crucial role in this process, which is tasked with encoding, storing, and recalling perceived information [2]. Thus, understanding the attributes that enhance memorability could unlock insights into perceptual mechanisms and guide the development of artificial systems that emulate human-like perception and memory retention.

The concept of memorability as defined in academic discourse is related to an inherent characteristic of an image or video that facilitates future recall [3]. Although subjective elements influence memory, recent research indicates that certain visual aspects consistently improve memorability [4–6], with studies revealing a link between the contextual and semantic properties of an image and its recall potential [7]. The analysis of image semantics extends to textual descriptions, providing a broader understanding of visual data [8].

The consideration of memorability in visual content reveals distinctions in memory retention; dynamic scenes and human presence often enhance memorability compared to static landscapes (Fig. 1). Leveraging these insights can significantly improve the effectiveness of audiovisual communication, enabling automatic systems to better capture audience attention and make memorable impacts.

(a) "A woman is water skiing off a boat by the lake" - Memorability: 0.86 (Highest tertile)

(b) "A pan view of a lake of ice during a cold day" - Memorability: 0.52 (Lowest tertile)

Fig. 1 These examples, sourced from the Memento10k dataset [9], illustrate the range of memorability within visual content. The memorability score, ranging from 0 to 1, predicts the probability of subsequent recall. a) This image exemplifies the association between dynamic settings and memorability. b) The lack of memorable features of the scene is attributed to its static and nondescript nature

While the studies mentioned above highlight the significant impact of visual and textual semantics on how well content is remembered, developing predictive systems to estimate multimedia memorability based on these semantic traits continues to be a challenging endeavor. This study aims to explore how various adaptation strategies can utilize these complex relationships to enhance the predictive performance of generalist pre-trained models. The advent of Contrastive Language Image Pre-training (CLIP) [10] presents a formidable approach in this context. CLIP facilitates the creation of shared embedding spaces that encapsulate the complex characteristics of both images and text. It offers a distinct advantage by jointly considering both sources: frames extracted from the clips and humanly annotated textual descriptions.

Several studies have explored the use of CLIP-based encoders as feature extractors for memorability prediction [11–15]. However, there is still no systematic evaluation that isolates the contributions of CLIP models and rigorously compares them to unimodal encoders trained only on visual or textual tasks. Open questions remain regarding the role of the text encoder, which captures multimodal semantic information but has been largely underexplored for memorability prediction. It is also unclear whether the performance improvements associated with CLIP arise purely from contrastive learning or from other factors such as model architecture. If contrastive learning enhances semantic representations in a way that benefits memorability prediction, additional adaptation to the specific domain of interest may further improve performance. To address these gaps, this work presents a comprehensive experimental study comparing CLIP-based image and text encoders with unimodal models under controlled conditions, ensuring similar architecture and parameter count across all configurations. The evaluation includes both conventional fine tuning, where model weights are adapted to the memorability task, and an additional contrastive pre-training stage using frame–text pairs from a memorability-annotated dataset. This stage, referred to as Fine tuned CLIP (FCLIP), investigates whether enriching CLIP encoders with task-specific semantic information leads to improved performance. Both adaptation strategies can be combined to maximize encoder alignment with the memorability prediction task. Through this framework, we aim to provide a rigorous comparison that clarifies the benefits of multimodal representations, contrastive learning, and domain adaptation for predicting video memorability.

In order to thoroughly study the effects of these adaptation steps, we develop different experimentation **strategies**:

1. **Unimodal Encoders.** Pre-trained encoders are utilized solely as feature extractors for constructing a regression model to predict memorability scores.
2. **Adapted Unimodal Encoders.** The same pre-trained encoders are fine tuned for the regression task, this time with their parameters unfrozen.
3. **CLIP-Based Encoders.** In this case, a pair of image and text encoders that have undergone contrastive pre-training using generic data pairs are used, potentially resulting in enhanced performance. It is worth noting that the relevant differences between these models and their unimodal counterparts (Strategies 1 and 2) do not lie on their architecture or size, but rather on their pre-training schema. We initially adopt these CLIP-based pre-trained encoders as feature extractors, similar to Strategy 1.
4. **Adapted CLIP-Based Encoders** As in Strategy 2, here we fine tune the original CLIP encoders by unfreezing their parameters.

5. **FCLIP-Based Encoders.** In this approach, instead of relying on a nonadapted pre-trained CLIP model we further fine tune that CLIP model using task-related data. Then, the resulting encoders, which have been adapted for memorability prediction in an unsupervised manner, are utilized as feature extractors. Consequently, the encoders remain frozen when linked to the regression model, which we train for the task.
6. **Fully Adapted FCLIP Encoders.** As in Strategies 2 and 4, here the parameters of the FCLIP model are set to evolve when learning the final task. Hence, they undergo two adaptation processes: first, through the unsupervised adaptation via CLIP model fine tuning, and second, through the supervised adaptation during regression model training for memorability prediction.

These strategies enable a systematic investigation of the role of multimodal pre-training, task-specific adaptation, and their combination in improving internal representations for memorability prediction. By enforcing consistent encoder architectures and regression model configurations across all strategies, we ensure a fair and controlled evaluation. In summary, the main contributions of this work are as follows:

1. We provide a comprehensive comparison between unimodal encoders and CLIP-based encoders, evaluating their ability to capture information relevant to memorability prediction under both feature extraction and fine tuning regimes.
2. We analyze the benefits of contrastive language-image pre-training for memorability prediction, isolating its impact relative to unimodal models with equivalent architecture and size.
3. We investigate the effect of FCLIP, a domain-adapted extension of CLIP obtained through additional contrastive pre-training on memorability-related data, and assess its contribution to improving encoder representations.
4. We evaluate the combined effect of unsupervised domain adaptation (via FCLIP) and supervised task-specific fine tuning, providing insights into the effectiveness of full encoder adaptation for memorability prediction.

Through this experimental framework, we aim to deliver a rigorous and comprehensive assessment of the most effective strategies for training and adapting encoders to predict video memorability.

The rest of this paper is structured as follows. Section 2 reviews the evolution of image and video memorability prediction, highlighting key advancements, limitations of existing CLIP-based approaches, and the motivation for this work.

Section 3 details our proposed approach, including the adaptation of CLIP encoders through an additional contrastive pre-training step and the methodology for integrating frames and textual descriptions for memorability prediction. The experimental setup, including the description of the data set, the pre-processing steps and the details of the training of the model, is presented in Section 4. Section 5 discusses the results of our experiments, providing insight into the effectiveness of our proposed approach and the comparison with the baseline methods. Finally, Section 6 concludes the paper with a summary of our findings and directions for future work.

2 Related work

The field of research has witnessed a surge of interest in the area of computational prediction of image and video memorability. Seminal studies in the neuroscience realm have provided the foundation for the development of automatic systems, which have been equipped with a diverse array of descriptors. These descriptors encompass a wide range of information, including low-level pixel data, scene and object information, and advanced semantic features. The advent of deep learning-based models, particularly cutting-edge vision-language models (VLMs), has led to substantial advancements in the domain.

2.1 Neuroscience and cognitive foundations

The processes used by the brain to remember visual information have been a topic of investigation since the late 1960s and early 1970s, primarily within the realms of psychology and neuroscience [16, 17]. Jaegle et al. [18] demonstrated the presence of different areas of the brain specialized in processing information related to memorability. Furthermore, Konkle et al. [19, 20] investigated the role of image-specific characteristics, such as conceptual structure, scene, and object representations, in relation to this cognitive factor. It should be noted that there exists an inverse relationship between human perception and the real memorability values, suggesting that this is indeed a counterintuitive notion [3].

2.2 Evolution of image memorability prediction

2.2.1 Low-level and scene-based features

Early computational efforts in image memorability prediction centered on low-level visual features. Isola et al. [3, 4] introduced the SUNMem dataset, showing that basic image statistics and object counts influence memorability. Their work highlighted connections between scene content and memorability, laying the groundwork for subsequent deep learning approaches.

2.2.2 Deep learning and attention mechanisms

With the advent of deep neural networks, Khosla et al. [21] fine tuned a convolutional model, MemNet, achieving significant performance gains over traditional descriptors. AMNet [22] introduced attention mechanisms and LSTM networks to iteratively refine predictions based on salient regions. EMNet [23] further incorporated emotional cues and saliency via an ensemble of CNNs, acknowledging the multifaceted nature of memorability.

2.2.3 Multimodal and CLIP-based advances

More recently, multimodal approaches that combine pre-trained visual and textual models like CLIP have emerged, with Multimodal Large Language Models (MM-LLMs) being a promising newcomer. Henry [24] combined visual CLIP features with textual, interaction, and behavioral data by means of the Llama13B LLM, achieving state-of-the-art results. However, most existing works either focus on frozen CLIP features [11, 12] or lack sys-

tematic comparison with unimodal baselines. The present study addresses this gap by evaluating both unimodal and CLIP-based encoders under controlled conditions, introducing additional domain adaptation through contrastive fine tuning (FCLIP). Table 1 provides an overview of image memorability predictors, highlighting the progression from handcrafted features to semantically enriched, multimodal models.

2.3 Video memorability prediction and multimodal advances

2.3.1 From visual cues to multimodal learning

Predicting memorability in video content extends image-based principles to dynamic, multimodal stimuli. The VideoMem [25] and Memento10k [9] datasets, collected via memory game protocols, have become benchmarks for this task. Initial models relied on visual cues alone. Cohendet et al. [25] applied semantic ResNet embeddings, while SemanticMemNet [9] fused visual and textual data with LSTM-based captioning, illustrating the benefits of multimodal signals.

Table 1 Overview of relevant works in image memorability prediction. Spearman Rank Correlation Coefficient (SRCC) results are indicative, experimental setups vary across systems. Abbreviations: FT = Fine tuned, FR = Frozen, SVR = Support Vector Regressor, SRCC = Spearman Rank Correlation Coefficient

Name	Input Features	Regressor	SRCC
SUNMem dataset			
Isola et al. [3]	Global visual features; Objects; Scene categories	SVR	0.540
Isola et al. [4]	Features from Isola 1; Spatial, content and aesthetic attributes	SVR	0.554
MemNet [21]	Raw images	FT Hybrid-CNN	0.630
AMNet [22]	Raw images; Attention maps from FT ResNet50	LSTM + Fully Connected	0.649
EMNet [23]	Raw images; Salient image patches	Ensemble of memorability and emotion-based FT VGG-16	0.664
Henry [24]	FR CLIP features from images; Textual, interaction, behavior factors; Metadata	FT LLaMa-13B	0.760
LaMem dataset			
MemNet [21]	Raw images	FT Hybrid-CNN	0.640
AMNet [22]	Raw images; Attention maps from FT ResNet50	LSTM + Fully Connected	0.677
EMNet [23]	Raw images; Salient image patches	Ensemble of memorability and emotion-based FT VGG-16	0.671
Henry [24]	FR CLIP features from images; Textual, interaction, behavior factors; Metadata	FT LLaMa-13B	0.720

2.3.2 CLIP integration and limitations

Kleinlein et al. [8] and Agarla et al. [11] employed CLIP-based encoders, leveraging frozen features for memorability regression. These approaches improved semantic understanding but did not explore encoder adaptation or contrastive domain fine tuning. Subsequent works incorporated scene layout [26], emotional cues [23], and behavioral data [27], underscoring the importance of high-level, context-aware representations. Recent studies also demonstrated that textual descriptions generated via ClipCap [29] or transformer-based vision-language models enhance memorability prediction.

Despite these advances, systematic evaluation of the relative contributions of multimodal pre-training, supervised fine tuning, and task-specific contrastive adaptation remains limited. The proposed work introduces FCLIP, which supplements generic CLIP encoders with additional domain-specific contrastive learning, and rigorously compares this approach to both unimodal and conventional CLIP baselines. Table 2 summarizes video memorability methods, illustrating performance trends across datasets and the increasing role of semantic, contextual, and multimodal features.

While CLIP encoders have demonstrated strong semantic representation capabilities, existing studies primarily utilize frozen models or shallow adaptation strategies. The role of further contrastive fine tuning with memorability-specific data (FCLIP) has not been systematically explored. Moreover, comparisons with unimodal models under matched architectural and parameter constraints are scarce. This work addresses these gaps by:

- Providing a controlled, comprehensive comparison of unimodal and CLIP-based encoders for memorability prediction.
- Benchmarking FCLIP, a domain-adapted extension of CLIP through contrastive pre-training with memorability-labeled data.
- Evaluating the effects of encoder adaptation, both unsupervised (contrastive) and supervised (task-specific fine tuning), on model performance.

These contributions offer new insights into the strengths, limitations, and optimization of multimodal encoders for video memorability prediction, advancing the field toward more effective, semantically grounded predictive systems.

3 Framework proposal

This section describes the proposed methodology for video memorability prediction, including the problem formulation, contrastive adaptation process, and supervised fine-tuning strategies applied to visual and textual encoders.

3.1 Problem statement

The task of multimodal video memorability prediction is formulated as a regression problem, where the objective is to predict a memorability score for each video sample. Each sample consists of a video and its associated textual descriptions. Each video is represented by N_{frames} key frames and N_{captions} textual captions, so that each sample is defined as a tuple

Table 2 Overview of relevant works in video memorability prediction. Spearman Rank Correlation Coefficient (SRCC) results are indicative, experimental setups vary across systems. Abbreviations: FT = Fine tuned, FR = Frozen, DNet = DenseNet-121, CLIP = CLIP model, SBERT = Sentence BERT, ST Attn = Spatio-Temporal Attention, SRCC = Spearman Rank Correlation Coefficient

Name	Input Features	Regressor	SRCC
VideoMem dataset			
Cohendet et al. [25]	Raw frames	FT semantic ResNet	0.503
SemanticMemNet [9]	FT DNet (frames); FT I3D (video, optical flow)	Late fusion; LSTM captioning	0.556
Kleinlein et al. [8]	FT DNet (frames); FR SBERT (captions)	Linear Regressor	0.450
Agarla et al. [11]	FR CLIP (frames)	MLP	0.470[1]
M3S [26]	Low-level descriptors; HRNetV2 (scene); CSN (events); Context features	MLP	0.563
Kumar et al. [14]	FT ResNet-CLIP (frames); ST Attn	MLP	0.505
Henry [24]	FR CLIP (frames); Textual, interaction, behavioral factors, metadata	FT LLaMa-13B	0.640
Behavior-LLaVa [27]	Raw video; Text captions, ASR, scene descriptions, metadata	FT LLaMa-Vid	0.600
Memento10k dataset			
SemanticMemNet [9]	FT DNet (frames); FT I3D (video, optical flow)	Late fusion; LSTM captioning	0.663
Sweeney et al. [28]	FR DNet (frames)	Bayesian Ridge	0.523
Kleinlein et al. [8]	FT DNet (frames); FR SBERT (captions)	Linear Regressor	0.600
Agarla et al. [11]	FR CLIP (frames)	MLP	0.523[1]
Guinandeau and Xalabarder [29]	FR SBERT (descriptions, captions); FR ResNet, DNet (frames)	MLP	0.629
Kleinlein et al. [12]	FR CLIP, ViT, BEIT (frames)	Bayesian Ridge	0.656
M3S [26]	Low-level descriptors; HRNetV2 (scene); CSN (events); Context features	MLP	0.670
Kumar et al. [14]	FT ResNet-CLIP (frames); ST Attn	MLP	0.713
Henry [24]	FR CLIP (frames); Textual, interaction, behavioral factors, metadata	FT LLaMa-13B	0.750
Behavior-LLaVa [27]	Raw video; Text captions, ASR, scene descriptions, metadata	FT LLaMa-Vid	0.710
LAMBDA dataset			
Henry [24]	FR CLIP (frames); Textual, interaction, behavioral factors, metadata	FT LLaMa-13B	0.550
Behavior-LLaVa [27]	Raw video; Text captions, ASR, scene descriptions, metadata	FT LLaMa-Vid	0.520

[1] Cross-dataset evaluation

(v, t), where $\mathrm{v} = (v_1, \ldots, v_{N_{\text{frames}}})$ are the selected frames and $\mathrm{t} = (t_1, \ldots, t_{N_{\text{captions}}})$ are the captions.

A model is trained to map each input x (either a frame v_i or a caption t_j) to a memorability score $\hat{y} \in [0, 1]$. This is achieved by first encoding the input using either a visual or textual encoder, f_{enc}^{M} $(M \in \mathrm{V}, \mathrm{T})$, to produce an embedding $z_M \in \mathbb{R}^d$. The embedding is then passed through a regression head g_{RH} to obtain the predicted score:

$$z_M = f_{\text{enc}}^{M}(x), \quad \hat{y} = g_{\text{RH}}(z_M) \tag{1}$$

To obtain a single prediction for the entire video sample, we adopt a late fusion approach. Specifically, we compute the memorability score for each frame or caption independently, and then aggregate these predictions by averaging:

$$\hat{y}_{\text{video}} = \frac{1}{N_{\text{frames}}} \sum_{i=1}^{N_{\text{frames}}} g_{\text{RH}}(f_{\text{enc}}^{V}(v_i)) \tag{2}$$

$$\hat{y}_{\text{text}} = \frac{1}{N_{\text{captions}}} \sum_{j=1}^{N_{\text{captions}}} g_{\text{RH}}(f_{\text{enc}}^{T}(t_j)) \tag{3}$$

In this work, we systematically analyze the effect of different adaptation strategies for the encoders, which are summarized as follows:

- **Contrastive adaptation:** Both visual and textual encoders are jointly fine tuned on image-text pairs from the memorability dataset by minimizing a contrastive loss. This step is formalized as a transformation $h_{\text{contrastive}}$ applied to the original encoders, producing adapted encoders $f_{\text{enc}}^{M,\text{FCLIP}} = h_{\text{contrastive}}(f_{\text{enc}}^{V}, f_{\text{enc}}^{T})$.
- **Supervised fine tuning:** The encoder and regression head are trained together on the memorability prediction task, minimizing a regression loss between the predicted and ground-truth scores. This can be formalized as $f_{\text{enc}}^{M,\text{SUP}} = h_{\text{supervised}}(f_{\text{enc}}^{M})$.

As a baseline, we also evaluate training only the regression head while keeping the encoder fixed. This approach can be applied to out-of-the-box unimodal or CLIP encoders, as well as to our FCLIP variants. By doing so, we assess the predictive power of the pre-trained features without further adaptation. A visual overview of this framework is provided in Fig. 2, and a summary of the combinations explored, described in Section 1, is shown in Table 3. The following sections describe each adaptation strategy and their mathematical formulation in detail.

3.2 Contrastive adaptation

In the contrastive adaptation step, the problem formulation shifts from direct memorability prediction to learning a shared representation for visual and textual inputs. The goal is to align the embedding spaces of the two modalities such that paired frames and captions from the same video are close in the representation space, while unpaired examples are pushed apart.

Formally, let $\mathcal{D}_{\text{pair}} = \{(v_i, t_i)\}_{i=1}^{N}$ denote a dataset of N image-text pairs, where each pair (v_i, t_i) corresponds to frames and captions from the same video. For this stage, we discard the memorability labels and focus solely on constructing these pairs: positive pairs are those sampled from the same video, and negative pairs are those sampled from different videos. To avoid ambiguity, we ensure that each mini-batch contains at most one frame and one caption per video, so there is only one correct pair per video in each batch.

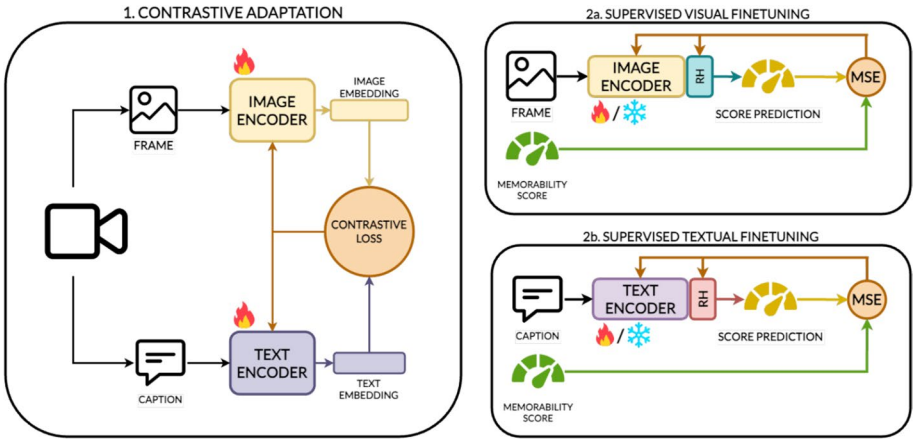

Fig. 2 The suggested pipeline for the two-phase fine tuning process of image and text encoders. In the first stage, the image and text encoders go through an additional pre-training phase within the contrastive learning framework, utilizing the frames and captions from the Memento10k dataset [9]. Moving on to the second stage, one of the encoders, either for text or image, is connected to a Regression Head (RH) and fine tuned for the memorability task using the provided dataset labels. This fine tuning process aims to minimize the Mean Square Error (MSE) between the predicted values and the actual memorability scores. The snowflake and flame symbols are used to indicate freezing and unfreezing of the model parameters, respectively. In the initial stage, the encoders are always adjusted, whereas in the subsequent stage, both scenarios are examined for their effectiveness

Table 3 Summary of the different strategies for adapting image and text encoders for memorability prediction. If contrastive adaptation is performed, the encoders are fine tuned using the CLIP schema using memorability related data (see Fig. 2). For the supervised fine tuning step, a snowflake means that the encoder weights are frozen, whilst a flame indicates that the encoder parameters are fine tuned when learning the regressor

Strategy	Pre-trained Checkpoint	Contrastive Adaptation	Supervised Fine Tuning
1 - Unimodal	`vit-base-patch32-224-in21kdistil-bert-base-uncased`	No	❄️
2 - Adapted Unimodal	`vit-base-patch32-224-in21kdistil-bert-base-uncased`	No	🔥
3 - CLIP-Based	`clip-vit-base-patch32`	No	❄️
4 - Adapted CLIP-Based	`clip-vit-base-patch32`	No	🔥
5 - FCLIP-Based	`clip-vit-base-patch32`	Yes	❄️
6 - Adapted FCLIP-Based	`clip-vit-base-patch32`	Yes	🔥

Given a batch of B image-text pairs $\{(v_k, t_k)\}_{k=1}^{B}$, we encode each image and text sample using the visual and textual encoders:

$$z_k^V = f_{\text{enc}}^V(v_k), \quad z_k^T = f_{\text{enc}}^T(t_k) \tag{4}$$

where $z_k^V, z_k^T \in \mathbb{R}^d$ are the respective modality embeddings.

The objective is to maximize the similarity between embeddings of matched pairs and minimize it for mismatched pairs. We use a contrastive loss based on Binary Cross-Entropy (BCE), computed for both modalities. For each image v_k, the ground truth is that t_k is the only positive caption and all others are negatives. The same holds when conditioning on each caption. The loss for a mini-batch is:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2B} \sum_{k=1}^{B} \left(\mathcal{L}_{\text{BCE}}(\mathrm{s}_k^{\text{img}}, \mathrm{y}_k) + \mathcal{L}_{\text{BCE}}(\mathrm{s}_k^{\text{txt}}, \mathrm{y}_k) \right) \tag{5}$$

where $\mathrm{s}_k^{\text{img}} = [\text{sim}(z_k^V, z_1^T), \ldots, \text{sim}(z_k^V, z_B^T)]$ is the similarity of image v_k to all captions in the batch, and $\mathrm{s}_k^{\text{txt}} = [\text{sim}(z_1^V, z_k^T), \ldots, \text{sim}(z_B^V, z_k^T)]$ is the similarity of caption t_k to all images. The label vector y_k is a one-hot vector indicating the correct match. The similarity function $\text{sim}(\cdot, \cdot)$ is implemented as the cosine similarity between normalized embeddings. The BCE loss for a batch of N predictions $\hat{y}_i$ and binary labels $y_i \in 0, 1$ is defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] \tag{6}$$

The encoders are adapted by minimizing this loss:

$$(f_{\text{enc}}^{V,\text{FCLIP}}, f_{\text{enc}}^{T,\text{FCLIP}}) = h_{\text{contrastive}}(f_{\text{enc}}^V, f_{\text{enc}}^T) = \arg \min_{f_{\text{enc}}^V, f_{\text{enc}}^T} \mathcal{L}_{\text{contrastive}} \tag{7}$$

After this adaptation, the resulting FCLIP encoders are better aligned to the semantics of the memorability dataset and can be used as initialization for subsequent supervised fine tuning.

3.3 Supervised fine tuning

In supervised fine tuning, the encoder (either visual or textual) and the RH are trained together to directly predict memorability scores. This is achieved by minimizing a regression loss, in this case the Mean Squared Error (MSE), between the predicted and ground-truth memorability scores. The adaptation process can be described as:

$$(f_{\text{enc}}^{M,\text{SUP}}, g_{\text{RH}}^{\text{SUP}}) = h_{\text{supervised}}(f_{\text{enc}}^M, g_{\text{RH}}) = \arg \min_{f_{\text{enc}}^M, g_{\text{RH}}} \mathcal{L}_{\text{regression}} \tag{8}$$

Where $\mathcal{L}_{\text{regression}}$ is the MSE loss for a batch of N predictions $\hat{y}_i$ and continuous labels y_i, given by:

$$\mathcal{L}_{\text{regression}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{9}$$

Alternatively, in the RH only setting, the encoder is kept frozen (i.e., $f\text{enc}^M$ is not updated), and only g_{RH} is optimized:

$$g_{\mathrm{RH}}^{\mathrm{frozen}} = \arg\min_{g_{\mathrm{RH}}} \mathcal{L}_{\mathrm{regression}} \quad \text{with } f_{\mathrm{enc}}^{M} \text{ frozen} \tag{10}$$

This allows us to disentangle the contribution of encoder adaptation from the representational power of the pre-trained features alone.

Inspired by Dong et al. [30], the structure of the RH is comprised of a Layer Normalization Module and a Linear Layer that projects the output vector to a single score prediction, along with a sigmoid layer to constrain the prediction within the range [0, 1]:

$$g_{\mathrm{RH}}(z) = \sigma\left(\mathrm{w}^{\top}\,\mathrm{LN}(z) + b\right) \tag{11}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization, w and b are the parameters of the linear layer, and $\sigma(\cdot)$ is the sigmoid activation function:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{12}$$

This lightweight approach enables the prediction of one-dimensional scores from the features extracted by the input encoder, allowing us to directly examine the relationship between these learned representations and memorability. In this way, we can assess how the studied adaptation strategies translate into downstream performance on this task.

3.4 Unimodal baseline

In order to test the effectiveness of the supervised adaptation step, we set up a baseline of unimodal encoders that are as similar as possible to the aforementioned CLIP encoders in terms of architecture, structure, and size, but have not been pre-trained using contrastive methods. For the vision branch, we resort to the ViT implementation that matches the one proposed in the CLIP paper, namely the `vit-base-patch32-224-in21k` [10, 31]. As for the textual branch, a thorough literature review did not yield any pretrained text Transformer model with the same architecture (i.e. the same number of encoders and hidden state dimension) as the CLIP encoder. For that reason, we choose the closest Transformer-based Language Model in terms of number of parameters, under the hypothesis that a relatively similar model in terms of general structure and size would be the most appropriate for fair comparison. The selected model is DistilBERT, a distilled version of the original BERT base model [32, 33]. In particular, the `distilbert-base-uncased` checkpoint is formed by 6 Transformer encoders with a hidden state dimension of 768, whilst the CLIP text model is composed by 12 Transformer encoders with a hidden state dimension of 512, accounting each for a total of 66M and 63M parameters, respectively.

4 Experimental setup

This section details the dataset, pre-processing steps, model configurations, and training procedures used to evaluate the proposed adaptation strategies under controlled conditions.

4.1 Dataset

In this study, the Memento10k dataset is used for experimentation [9]. The dataset was introduced in 2020 by Newman et al. and consists of 10,000 brief, 3 second long videos. Its purpose is to analyze the decrease in video recall based on the time gap between repetitions in the memory game mentioned before. Given the short length of the videos, it can be assumed that a single semantic unit is represented. One distinctive feature of the dataset is that the videos are sourced from web scraping and typically have lower image quality due to being recorded with affordable consumer equipment. However, the focus is on human actions and movement, with the samples displaying a significant amount of optical flow.

In the specific memory game designed to curate this corpus, human annotators sourced using Amazon Mechanical Turk were presented a series of target clips ranging from 9 videos (less than 30 seconds) to 200 videos (about 9 minutes). The game was played in a single session. After conducting a qualitative analysis of the annotations, the researchers found that films containing individuals, facial expressions, hands, simulated settings, and dynamic objects tended to be more memorable compared to those featuring outdoor landscapes or scenes that were somber, chaotic, or unchanging. The authors analyze the raw scores, which represent the percentage of correct identifications of the video after review, to adjust for the different delays in the presentation of the samples in the target clips T.

The Memento10k dataset contains five text descriptions for each video, known as Closed Captions (CC) or captions. These textual descriptions effectively summarize the video characteristics in a clear manner, focusing on semantic details and omitting emotional elements. This aspect encourages the application of bimodal learning approaches, which merge textual information with visual features and contextual significance to generate representations.

4.1.1 Data splits - contrastive adaptation

The MediaEval Predicting Video Memorability task provides an official train-validation-test data split for the Memento10k dataset. Annotated scores are available for the 7,000 train and 1,500 validation videos, but not for the 1,500 test clips. In the contrastive adaptation step, both the official training and validation splits are used for model adaptation. As this step is self-supervised and does not require label annotations, the 1,500 samples in the corpus test set are used to monitor the out of distribution loss for this pre-training step, in order to make an informed decision on which checkpoint to use for the latter stages. The initial, middle and last frames of each clip are extracted for training, resulting in $8,500 * 3 = 25,500$ training images. In addition, every chosen video frame is matched with each of the five textual descriptions provided, leading to a total of $25,500 * 5 = 127,500$ pairs of training images and captions. The same is done for the testing set, resulting in $1,500 * 3 * 5 = 22,500$ sample pairs for evaluation.

4.1.2 Data splits - supervised fine tuning

At this stage, the same frame extraction process is repeated. We use a K-Fold cross-validation (CV) scheme with $K = 5$ on the train and validation sets in order to assess the performance of the proposed approaches. Hence, for each fold a grand total of $8,500 * \frac{4}{5} * 3 = 20,400$ images or $8,500 * \frac{4}{5} * 5 = 34,000$ textual captions are used for training. As explained ear-

lier, to generate a unified prediction for the entire video during evaluation, the average of the predictions for each frame or caption is calculated using a late fusion strategy. Therefore, although $8,500 * \frac{1}{5} * 3 = 5,100$ video frames or $8,500 * \frac{1}{5} * 5 = 8,500$ captions are used for inference, the evaluation metrics are computed for the fused prediction scores, which add up to $8,500 * \frac{1}{5} = 1,700$ samples.

4.2 Image and text processing

During the pre-training step, the images and captions are pre-processed using the transformations defined by the original CLIP implementation. For the visual branch, the frames are center-cropped to 224x224 pixels and RGB normalized. With respect to the text, the CLIP Byte Pair Encoding (BPE)-based pre-trained tokenizer is used.

4.3 Adapting CLIP pre-training

During the pre-training phase, the encoders are adjusted to reduce the Contrastive Loss function by employing the Adam optimizer, with a batch size of 32 samples. The initial learning rate is automatically determined using the method described by Smith [34]. This involves performing multiple training iterations with different learning rates and monitoring the corresponding loss values. The selected specific learning rate is the one at which the loss is placed at the midpoint of the steepest descent region on the graph that shows the relationship between learning rate and loss. We set the learning rate bounds at $[10^{-8}, 1]$ and find an optimum value of $2.75 * 10^{-4}$. The authors provide empirical evidence to demonstrate the consistent effectiveness of this approach. To avoid overshooting, a "Reduce on Plateau" learning rate scheduler is used, reducing the value by a factor of ten each time the validation loss does not increase for three consecutive epochs. Furthermore, an early stopping strategy is employed to prevent overfitting. Training is halted if no improvement in test loss is observed for six consecutive epochs. A checkpoint is saved at the end of each epoch, and the model with the lowest contrastive loss on the test set is selected for subsequent memorability fine-tuning.

4.3.1 Fine tuning image and text encoders

At this stage, the models are also trained using the Adam optimizer with an automatically selected learning rate, applying the same procedure as in the contrastive adaptation step. The batch size hyperparameter is validated using a K-Fold CV scheme with $K = 5$, and the best configuration is reported in terms of the mean evaluation metric. Most importantly, the splits are computed at the video level, ensuring that there is no data leakage in the form of frames or captions belonging to the same clip appearing in the training and validation sets at the same time. The range of batch sizes explored during hyperparameter tuning depends on both the modality of the encoder (image or text) and whether the encoder is frozen or fine-tuned during training. When the encoder is frozen and only the regression head is trained, the number of trainable parameters and the computational burden are greatly reduced. This allows for significantly larger batch sizes, as memory and compute resources are not consumed by backpropagation through the encoder. In contrast, fine tuning the encoder increases both the number of trainable parameters and the memory requirements, which

limits the maximum feasible batch size. This distinction is reflected in the batch size bounds reported in Table 4.

Training is conducted for up to 200 epochs, with early stopping applied to prevent overfitting, using validation Spearman Correlation as a monitored metric. If the validation metric does not improve for 20 consecutive epochs after the 50^{th} epoch, training is halted. We report the performance obtained by the best checkpoint in terms of SRCC.

All experiments in both stages (contrastive and supervised adaptation) are performed using a single NVIDIA® GeForce RTX™ 4090 24GB GPU.

4.3.2 Evaluation metric

In the video memorability prediction literature, the Spearman rank correlation coefficient (SRCC) is widely used for model evaluation and strategy comparison. The SRCC is a non-parametric measure of the monotonic relationship between predictions and targets. In contrast to other metrics, the Spearman correlation evaluates the trend and direction of data considering rank order, making it a useful tool for evaluating nonlinear relationships. The results of the experiments are reported in terms of the mean SRCC over the 5 folds with a 95% Confidence Interval (CI) that evaluates the statistical significance of the provided results. CI is computed using the Fisher z transformation as a function of the mean correlation over the folds *SRCC* and the number of evaluation samples $n = 8,500$ [35]:

$$\tanh(\text{atanh}(SRCC) \pm \frac{1,96}{\sqrt{n-3}}) \tag{13}$$

5 Results and discussion

This section presents the experimental results, comparing different encoder configurations and adaptation strategies, followed by a discussion of their implications for video memorability prediction.

5.1 Pre-training adaptation step

Table 5 shows the evolution of the Cross Entropy loss between predictions and target for the training data (in this case, the concatenation of the training and validation splits) and testing data (the official test split). It can be seen that training for a single epoch drastically decreases test loss, which points to the model being able to learn the similarities and differences between the texts and images in the corpus, obtaining encoders that are more semantically aware of the nuances of the kind of multimedia data involved in the task. However, although the training loss still decreases for the following epochs, it is not the case for the test loss, which increases in a clear sign of heavy overfitting. This outcome was somewhat

Table 4 Summary of the batch size ranges explored for each encoder modality (Image or Text) and training setting (Frozen or Fine Tuned)

Encoder Modality	Frozen	Fine Tuned
Image	256 - 2,048	64 - 512
Text	512 - 4,096	64 - 512

Table 5 Training and evaluation loss for the contrastive pre-training adaptation of image and text encoders. Epoch 0 corresponds to an evaluation pass using the default encoders on the testing set without adaptation

Epoch	Train Loss	Test Loss
0 (No training)	–	1.93658
1	0.2187	**0.4209**
2	0.0912	0.4538
3	0.0610	0.4814
4	0.0458	0.4858
5	0.0374	0.5736
6	**0.0316**	0.5710

Table 6 Results for the comparison study between the proposed strategies. FCLIP denotes our proposed encoders that undergo an additional contrastive pre-training step on Memento10k. The best result across all experiments for each modality (vision or text) is shown in bold

Strategy	Encoder Model	Learning Rate	Batch Size	SRCC (CI)
1 - Unimodal	ViT	2e-03	1,024	0.603 (0.589, 0.616)
	DistilBERT	1e-03	512	0.574 (0.559, 0.588)
2 - Adapted Unimodal	ViT	3e-05	128	0.622 (0.609, 0.635)
	DistilBERT	1e-05	256	0.629 (0.616, 0.641)
3 - CLIP-Based	CLIP - Vision	1e-03	1,024	0.664 (0.651, 0.675)
	CLIP - Text	1e-03	4,096	0.609 (0.595, 0.622)
4 - Adapted CLIP-Based	CLIP - Vision	2e-06	128	0.670 (0.658, 0.681)
	CLIP - Text	2e-06	256	0.619 (0.606, 0.632)
5 - FCLIP-Based	**FCLIP - Vision**	**1e-03**	**2,048**	**0.672 (0.660, 0.683)**
	FCLIP - Text	2e-03	512	0.628 (0.615, 0.640)
6 - Adapted FCLIP-Based	FCLIP - Vision	2e-06	128	0.671 (0.659, 0.682)
	FCLIP - Text	**3e-06**	**512**	**0.632 (0.619, 0.644)**

predictable since the original encoders were already positioned at a fairly optimal spot in the loss function due to their initial extensive pre-training. Consequently, subjecting them repeatedly to a relatively small set of samples and requiring them to adjust may cause them to deviate from their starting point and end up in less-than-ideal setups.

Notwithstanding this, it is reasonable to assume that the encoders, which have been adjusted for a single iteration through the data, have already been exposed to and absorbed the particular subtleties of the dataset. Therefore, the saved checkpoint at the end of the initial epoch, which minimizes validation loss, is employed for fine tuning.

5.2 Fine tuning encoders for memorability prediction

A comparative overview of the performance of each strategy in terms of SRCC is shown in Table 6. The evolution of these results across strategy families is further illustrated in Fig. 3.

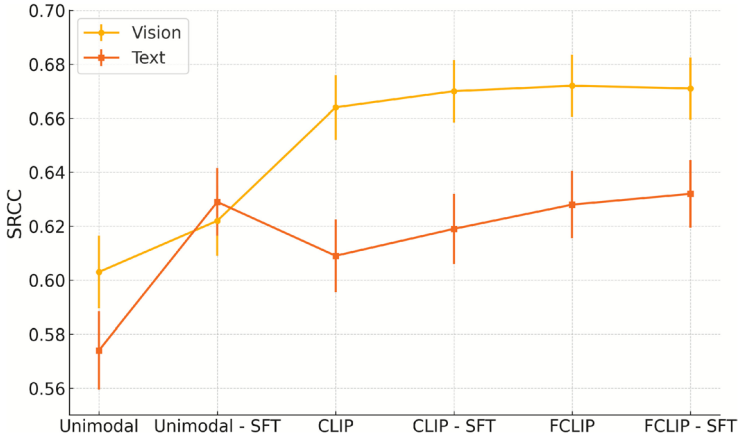

Fig. 3 Performance trend for each encoder strategy evaluated in terms of SRCC. The evolution of both visual and textual branches is shown, progressing from unimodal encoders through contrastively pre-trained CLIP encoders, to FCLIP encoders incorporating memorability-driven contrastive adaptation. The suffix-*SFT* denotes supervised fine tuning on memorability labels. Results highlight the consistent benefit of multimodal pre-training and the additional, though more modest, gains from task-specific fine tuning

The best results for both the visual and textual branch are obtained using FCLIP encoders, which have undergone a contrastive adaptation step (0.672 for vision and 0.632 for text). The gradual improvements leading to these results can be clearly observed in Fig. 3.

5.2.1 Visual encoder results and adaptation

Regarding visual encoders, results show that there is a significant improvement in performance when using a CLIP visual encoder instead of the unimodal ViT, both when freezing the model parameters (0.664 vs. 0.603) and when letting them evolve (0.670 vs. 0.622). This trend aligns with the progression observed in Fig. 3, where the transition from unimodal to CLIP-based encoders produces consistent performance gains. This points to the importance of incorporating semantic knowledge into image representations when predicting memorability, as combining information from the pixel values themselves with a higher level, comprehensive knowledge about the actions and elements appearing in the scene becomes crucial for the task at hand. Moreover, acquiring this knowledge through the use of textual descriptions seems adequate in light of the metrics obtained. FCLIP encoders slightly outperform their CLIP counterparts (0.672 vs. 0.664 when frozen), although their difference is not significant under this setup. This might be due to the fact that this second adaptation step is done with relatively fewer data points than the ones used on the original pre-training phase, which would make the encoders less sensitive to newly exposed samples. Furthermore, the amount of new information that the encoders learn throughout this process may be limited by employing a checkpoint that has only been trained for one epoch, which may encourage future research into ways to improve and modify the pre-training step tactics.

The results show that the supervised adaptation step results in a slight improvement delta in terms of SRCC for the ViT and CLIP encoders (0.622 vs. 0.603 and 0.670 vs. 0.664 respectively), but this is not the case for the FCLIP encoder (0.671 vs. 0.672). The plateauing trend in Fig. 3 further confirms this limited impact of supervised fine tuning at

the FCLIP stage. The original large-scale pre-training step generates robust enough representations that serve as powerful embeddings for learning this downstream task, so that further adapting them on a supervised manner does not translate to a significant boost in performance. There exists additional evidence on other supervised tasks of CLIP models not being able to stand out when fine tuned [36]. This motivates further exploration on efficient and effective strategies for CLIP model fine tuning with the aim of making the most of its vast semantic pre-training knowledge. Additionally, for this particular case, there might not be enough additional variability in the new data that are exposed to the models. In other words, the captions in Memento10k tend to be objective descriptions of what is seen in the video and sometimes lack subjective information that could be beneficial in order to link their representations to the memorability task. However, the results with respect to the unimodal image model are in line with the experiences reported on other computer vision problems [37–39].

All in all, the experiences reported thus far tend us to conclude that there is a significant gain on leveraging image encoders that have been extensively pre-trained on multimodal data using a contrastive schema such as CLIP. These models are able to capture semantic aspects of the images and relate them to their memorability potential. Contrastive adaptation of these image encoders using memorability related data provides an additional performance boost by learning the details and nuances of memorability related images and text, although a potential lack of additional variance in the newly presented data with respect of the original pre-training corpus makes this difference not significant. Lastly, a supervised adaptation of the encoders yields marginal improvements with respect to the same models that are used as feature extractors, which is also reflected in the flattened segments of the performance curve shown in Fig. 3.

5.2.2 Textual encoder results and adaptation

With respect to the textual branch, there is also a significant improvement when using frozen CLIP language encoders instead of a text-only pre-trained Transformer of similar characteristics (0.609 vs. 0.574). This shows that, in the same way that image encoders benefit from learning semantic information through text, language models are also subject to improvement when exposed to the pixel information that matches what is being described in the caption. There is also a performance gain when the contrastive fine tuning step is performed (0.628 vs. 0.609 for FCLIP/CLIP frozen and 0.632 vs. 0.619 for FCLIP/CLIP unfrozen), consistent with the upward trend for the textual branch observed in Fig. 3. Although the best result is obtained when undergoing both adaptation processes, contrastive and supervised (0.632), the second best comes from just fine tuning the unimodal DistilBERT Transformer using the memorability labels (0.629). It is worth noting that although the DistilBERT and CLIP-Text encoder stacks share the same number of parameters, the higher dimensionality of DistilBERT Transformer hidden states may suppose an advantage on the complexity of the information it can acquire when fine tuned. However, to the best of our knowledge, there is a lack of experimentation on the overall adequacy of the CLIP text encoder on downstream tasks, neither multimodal nor classic NLP-related. This motivates further work to investigate the full potential of this type of approach in predicting memorability from textual descriptions, as well as other downstream regression or classification tasks.

5.2.3 Hyperparameter selection and variability across experiments

The variability in optimal batch sizes and learning rates across experiments reflects the combined influence of encoder pre-training depth, parameter freeze status, and data modality. For instance, when using frozen encoders such as the CLIP vision model (Strategy 3), the optimal batch size reached 1,024 with a learning rate of 1e-3, leveraging the minimal gradient computation required when only the regression head is updated. In contrast, fully fine-tuned CLIP models (Strategy 4) and FCLIP models (Strategy 6) required significantly smaller learning rates, in the order of 2e-6 to 3e-6, and batch sizes reduced to 128 or 256, to prevent overfitting and to stabilize updates to semantically rich pre-trained weights. This effect is particularly evident in the vision branch, where FCLIP encoders achieved the best SRCC of 0.672 with a large batch size of 2,048 when frozen, but required much more conservative settings (BS = 128, LR = 2e-6) during full fine-tuning to avoid performance degradation. Similar trends were observed in the textual branch, where CLIP-Text encoders supported batch sizes up to 4,096 when frozen but required reductions to 256–512 during fine-tuning, illustrating how both computational constraints and model sensitivity to parameter updates dictate the observed hyperparameter differences.

6 Conclusions and future work

This work presents a systematic evaluation of pre-trained multimodal encoders for video memorability prediction, with a focus on contrastive learning, domain adaptation, and supervised fine tuning. We introduced FCLIP, an extension of CLIP encoders that undergoes additional contrastive pre-training on memorability-specific image-text pairs. Through controlled experiments, we compared FCLIP, standard CLIP models, and unimodal baselines to assess their effectiveness in predicting video memorability.

The results demonstrate that multimodal contrastive pre-training significantly enhances memorability prediction performance relative to unimodal encoders. FCLIP models, which incorporate task-specific adaptation, provide further - albeit modest - performance gains, underscoring the value of aligning semantic and visual information for this task.

6.1 Limitations

Despite these encouraging findings, several limitations should be acknowledged. First, the memorability-specific dataset used for FCLIP pre-training is comparatively small relative to the large-scale datasets employed during the original CLIP training. This constrains the degree to which models can effectively adapt to the domain and increases the risk of overfitting, as observed beyond the first epoch of contrastive adaptation. Furthermore, the CLIP models used in this study are inherently image-based, meaning they do not explicitly model temporal dependencies within video sequences. Memorability predictions are derived by averaging frame-wise outputs, a simplification that overlooks the dynamic nature of video content and the potential influence of temporal patterns on memorability. Additionally, the textual descriptions available in the Memento10k dataset are largely objective and factual, lacking subjective, emotional, or context-rich elements that may better capture human memory cues.

6.2 Future work directions

To address these limitations and further advance the field, several research directions are proposed:

- Exploring pre-training on larger, more diverse multimodal datasets such as LAMBDA [24] or Movie Memorability [40] to enhance the generalization and robustness of FCLIP models.
- Integrating temporal modeling approaches, such as temporal attention mechanisms or video-specific transformer architectures, to capture sequential dependencies and dynamics inherent in video memorability.
- Augmenting textual descriptions with subjective, synthetic captions that encode affective or perceptual information, thereby enriching the semantic grounding of the models.
- Investigating the potential of emerging multimodal LLMs and Vision-Language Models (VLMs) to enhance memorability prediction, while considering their domain adaptation requirements and data constraints.

6.3 Ethical considerations

The ability to predict and understand video memorability holds potential for positive societal impact. This includes fostering media literacy by elucidating the factors that influence content retention, supporting ethical marketing and educational content creation, and informing regulatory frameworks that promote responsible content dissemination.

However, this technology also presents risks. Predictive systems may be exploited to manipulate user attention, propagate deceptive or persuasive content, or reinforce cognitive biases. Future research should prioritize the development of mechanisms to assess and mitigate these risks, ensuring that advances in memorability prediction contribute to socially beneficial, ethically responsible applications.

Acknowledgements The research of Iván Martín-Fernández was supported by the Universidad Politécnica de Madrid (Programa Propio I+D+i).

Author contributions Iván Martín-Fernández: Conceptualization of this study, Software, Experimentation, Writing - Original draft, Writing - Final version.
Sergio Esteban-Romero: Software, Writing - Original draft.
Manuel Gil-Martín: Conceptualization of this study, Methodology, Writing - Final version.
Fernando Fernández-Martínez: Conceptualization of this study, Methodology, Writing — Final version, Funding.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was funded by Project ASTOUND (101071191 — HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) of the European Commission and by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C22), TRUSTBOOST (PID2023-150584OB-C21) and BeWord (PID2021-126061OB-C43), funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR".

Data availability This research uses previously released data. In particular, the Memento10k dataset releases labeled videos for their training and development set and unlabeled videos for their test set [9].

Declarations

Competing interests The authors report no competing interests relevant to the content of this article.

References

1. Bernstein DA, Nash PW (2008) Sensation and Perception, pp 84–134. Houghton Mifflin Company, Boston, MA
2. Bernstein DA, Nash PW (2008) Memory, pp. 207–245. Houghton Mifflin Company, Boston, MA
3. Isola P, Xiao J, Torralba A, Oliva A (2011) What makes an image memorable? In: CVPR 2011, pp 145–152. https://doi.org/10.1109/CVPR.2011.5995721
4. Isola P, Parikh D, Torralba A, Oliva A (2011) Understanding the intrinsic memorability of images. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. NIPS'11, pp 2429–2437. Curran Associates Inc., Red Hook, NY, USA. https://doi.org/10.5555/2986459.2986730
5. Lin Q, Yousif SR, Chun MM, Scholl BJ (2021) Visual memorability in the absence of semantic content. Cognition 212:104714. https://doi.org/10.1016/j.cognition.2021.104714
6. Xie W, Bainbridge WA, Inati SK, Baker CI, Zaghloul KA (2020) Memorability of words in arbitrary verbal associations modulates memory retrieval in the anterior temporal lobe. Nat Hum Behav 4(9):937–948. https://doi.org/10.1038/s41562-020-0901-2
7. Bylinskii Z, Goetschalckx L, Newman A, Oliva A (2022) Memorability: An image-computable measure of information utility. Human Perception of Visual Information: Psychological and Computational Perspectives, 207–239
8. Kleinlein R, Luna-Jiménez C, Arias-Cuadrado D, Ferreiros J, Fernández-Martínez F (2021) Topic-oriented text features can match visual deep models of video memorability. Appl Sci 11(16):7406. https://doi.org/10.3390/app11167406
9. 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: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16, pp 223–240. https://doi.org/10.1007/978-3-030-58517-4_14 . Springer
10. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning transferable visual models from natural language supervision. In: Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol 139, pp 8748–8763. PMLR, Virtual Event. https://proceedings.mlr.press/v139/radford21a.html
11. Agarla M, Celona L, Schettini R et al (2023) Predicting video memorability using a model pretrained with natural language supervision. In: MediaEval multimedia benchmark workshop working notes
12. Kleinlein R, Luna-Jiménez C, Fernández-Martínez F (2021) Thau-upm at mediaeval 2021: From video semantics to memorability using pretrained transformers. In: MediaEval multimedia benchmark workshop working notes
13. Martín-Fernández I, Kleinlein R, Luna-Jiménez C, Gil-Martín M, Fernández-Martínez F (2023) Video memorability prediction from jointly-learnt semantic and visual features. In: Proceedings of the 20th international conference on content-based multimedia indexing. CBMI '23, pp 178–182. Association for Computing Machinery, New York, NY, USA.https://doi.org/10.1145/3617233.3617260
14. Kumar P, Khandelwal E, Tapaswi M, Sreekumar V (2025) Seeing eye to ai: Comparing human gaze and model attention in video memorability. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV), pp 2082–2091. https://doi.org/10.1109/WACV61041.2025.00209
15. Sweeney L, Healy G, Smeaton AF (2022) Diffusing surrogate dreams of video scenes to predict video memorability. In: MediaEval multimedia benchmark workshop working notes

16. Shepard RN (1967) Recognition memory for words, sentences, and pictures. J Verbal Learn Verbal Behav 6(1):156–163. https://doi.org/10.1016/S0022-5371(67)80067-7
17. Standing L (1973) Learning 10000 pictures. Quarterly Journal of Experimental Psychology 25(2):207–222. https://doi.org/10.1080/14640747308400340
18. Jaegle A, Mehrpour V, Mohsenzadeh Y, Meyer T, Oliva A, Rust N (2019) Population response magnitude variation in inferotemporal cortex predicts image memorability. eLife 8, 47596. https://doi.org/10.7554/eLife.47596
19. Konkle T, Brady TF, Alvarez GA, Oliva A (2010) Scene memory is more detailed than you think: The role of categories in visual long-term memory. Psychol Sci 21(11):1551–1556. https://doi.org/10.1177/0956797610385359. PMID: 20921574
20. Konkle T, Brady TF, Alvarez GA, Oliva A (2010) Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. J Exp Psychol Gen 139(3):558
21. Khosla A, Raju AS, Torralba A, Oliva A (2015) Understanding and predicting image memorability at a large scale. In: 2015 IEEE international conference on computer vision (ICCV), pp 2390–2398 . https://doi.org/10.1109/ICCV.2015.275
22. Fajtl J, Argyriou V, Monekosso D, Remagnino P (2018) Amnet: Memorability estimation with attention. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 6363–6372. https://doi.org/10.1109/CVPR.2018.00666
23. Basavaraju S, Sur A (2019) Multiple instance learning based deep cnn for image memorability prediction. Multimedia Tools and Applications 78(24):35511–35535. https://doi.org/10.1007/s11042-019-08202-y
24. Si H, Singh S, Singla YK, Bhattacharyya A, Baths V, Chen C, Shah RR, Krishnamurthy B (2025) Long-term ad memorability: Understanding & generating memorable ads. In: Proceedings of the winter conference on applications of computer vision (WACV), pp 5707–5718. https://doi.org/10.1109/WACV61041.2025.00557
25. Cohendet R, Demarty C-H, Duong NQ, Engilberge M (2019) Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2531–2540. https://doi.org/10.1109/ICCV.2019.00262
26. Dumont T, Hevia JS, Fosco CL (2023) Modular memorability: Tiered representations for video memorability prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10751–10760. https://doi.org/10.1109/CVPR52729.2023.01035
27. Singh SK, I HS, Singla YK, Chen C, Shah RR, Baths V, Krishnamurthy B (2025) Teaching human behavior improves content understanding abilities of VLMs. In: The thirteenth international conference on learning representations. https://openreview.net/forum?id=ff2V3UR9sC
28. Sweeney L, Healy G, Smeaton AF (2021) Predicting media memorability: comparing visual, textual and auditory features. In: MediaEval multimedia benchmark workshop working notes
29. Guinaudeau C, Xalabarder AG (2023) Textual analysis for video memorability prediction. In: MediaEval multimedia benchmark workshop working notes (2023)
30. Dong X, Bao J, Zhang T, Chen D, Shuyang G, Zhang W, Yuan L, Chen D, Wen F, Yu N (2022) Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. arXiv preprint arXiv:2212.06138, https://doi.org/10.48550/arXiv.2212.06138
31. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: International conference on learning representations. https://openreview.net/forum?id=YicbFdNTTy
32. Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers), pp 4171–4186. https://doi.org/10.18653/v1/N19-1423
33. Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, https://doi.org/10.48550/arXiv.1910.01108
34. Smith LN (2017) Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision (WACV), pp 464–472. https://doi.org/10.1109/WACV.2017.58
35. Fisher RA (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10(4):507–521. Accessed 2024-05-21
36. Kumar A, Raghunathan A, Jones RM, Ma T, Liang P (2022) Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International conference on learning representations. https://openreview.net/forum?id=UYneFzXSJWh
37. Debenedetti E, Sehwag V, Mittal P (2023) A light recipe to train robust vision transformers. In: 2023 IEEE conference on secure and trustworthy machine learning (SaTML), pp 225–253. https://doi.org/10.1109/SaTML54575.2023.00024

38. Atito S, Awais M, Kittler J (2021) Sit: Self-supervised vision transformer. arXiv preprint arXiv:2104.03602, https://doi.org/10.48550/arXiv.2104.03602
39. Liu Y, Sangineto E, Bi W, Sebe N, Lepri B, Nadai MD (2021) Efficient training of visual transformers with small datasets. In: Beygelzimer A, Dauphin Y, Liang P, Vaughan JW (eds) Advances in Neural Information Processing Systems. https://openreview.net/forum?id=SCN8UaetXx
40. Cohendet R, Yadati K, Duong NQK, Demarty C-H (2018) Annotating, understanding, and predicting long-term video memorability. In: Proc of the ICMR 2018 Workshop, Yokohama, Japan, June 11-14 (2018). https://doi.org/10.1145/3206025.3206056

Authors and Affiliations

Iván Martín-Fernández[1] · Sergio Esteban-Romero[1] · Manuel Gil-Martín[1] · Fernando Fernández-Martínez[1]

✉ Iván Martín-Fernández
ivan.martinf@upm.es

Sergio Esteban-Romero
sergio.estebanro@upm.es

Manuel Gil-Martín
manuel.gilmartin@upm.es

Fernando Fernández-Martínez
fernando.fernandezm@upm.es

[1] 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), Av. Complutense, 30, Madrid 28040, Madrid, Spain