Video Memorability Prediction From Jointly-learnt Semantic and Visual Features

Iván Martín-Fernández
ivan.martinf@upm.es
THAU Group, IPTC
Universidad Politécnica de Madrid
Madrid, Spain

Ricardo Kleinlein
ricardo.kleinlein@upm.es
THAU Group, IPTC
Universidad Politécnica de Madrid
Madrid, Spain

Cristina Luna-Jiménez
cristina.lunaj@upm.es
THAU Group, IPTC
Universidad Politécnica de Madrid
Madrid, Spain

Manuel Gil-Martín
manuel.gilmartin@upm.es
THAU Group, IPTC
Universidad Politécnica de Madrid
Madrid, Spain

Fernando Fernández-Martínez
fernando.fernandezm@upm.es
THAU Group, IPTC
Universidad Politécnica de Madrid
Madrid, Spain

ABSTRACT

The memorability of a video is defined as an intrinsic property of its visual features that dictates the fraction of people who recall having watched it on a second viewing within a memory game. Still, unravelling what are the key features to predict memorability remains an obscure matter. This challenge is addressed here by fine-tuning text and image encoders using a cross-modal strategy known as Contrastive Language-Image Pre-training (CLIP). The resulting video-level data representations learned include semantics and topic-descriptive information as observed from both modalities, hence enhancing the predictive power of our algorithms. Our proposal achieves in the text domain a significantly greater Spearman Rank Correlation Coefficient (SRCC) than a default pre-trained text encoder (0.575 ± 0.007 and 0.538 ± 0.007, respectively) over the Memento10K dataset. A similar trend, although less pronounced, can be noticed in the visual domain. We believe these findings signal the potential benefits that cross-modal predictive systems can extract from being fine-tuned to the specific issue of media memorability.

CCS CONCEPTS

• **Information systems → Multimedia information systems**.

KEYWORDS

media memorability, CLIP, cross-modal, pre-training

ACM Reference Format:
Iván Martín-Fernández, Ricardo Kleinlein, Cristina Luna-Jiménez, Manuel Gil-Martín, and Fernando Fernández-Martínez. 2023. Video Memorability Prediction From Jointly-learnt Semantic and Visual Features. In *20th International Conference on Content-based Multimedia Indexing (CBMI 2023), September 20–22, 2023, Orléans, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3617233.3617260

CBMI 2023, September 20–22, 2023, Orléans, France

ACM ISBN 979-8-4007-0912-8/23/09.
https://doi.org/10.1145/3617233.3617260

1 INTRODUCTION

Identifying the features of a particular audiovisual communication medium that make it memorable is becoming increasingly relevant due to the expanding amount of available multimedia content. Memorability is defined in the literature as an intrinsic property of an image or a video associated with its easiness to be recalled in subsequent viewings, and that is a function of its audiovisual features [9]. Recent studies from psychology and neurosciences seem to disagree with the idea that memory is an entirely subjective appraisal, instead suggesting that there are indeed visual elements that are more likely to be stored in memory for later recall [8, 15, 25]. Memorability is an observer-independent aspect of the visual medium, greatly influenced by the semantics of the scenes it represents [3], which motivates the use of alternative sources to analyse it beyond the purely visual domain, for instance, employing text-based captions that describe a scene [12].

We hypothesise that these two modalities can be fused together in order to learn powerful content descriptors that take into consideration features learned from both text and visual domains. In turn, this would enhance their separate predictive capabilities when adapted to the task of media memorability. Consequently, in this paper we propose a method that jointly learns image and text-based features important to predict video memorability from video frames and video-level captions, fine-tuning Transformer-based encoders using a cross-modal, contrastive learning approach.

The rest of this paper is structured as follows: Section 2 provides an overview of related work; in Section 3, we present our proposal for video memorability prediction using jointly learnt semantic and visual features; Section 4 describes the experimental setup, including the Memento10K dataset employed throughout this study; afterwards, Section 5 introduces and discusses the results obtained; finally, in Section 6, we draw the main conclusions of our work and outline potential open avenues for future work.

2 RELATED WORK

The Transformer architecture has marked an important milestone towards pushing forward the state-of-the-art across several downstream tasks [24]. This family of models has shown a remarkable ability to build robust, semantically-rich embedding features from input sources of different domains, such as the Vision Transformer

(ViT) for images [6] or BERT for text [5]. This makes them particularly appealing for blending modalities. One scheme that is frequently used to jointly learn the aforementioned representations (merging image and text-based information), is Contrastive Language-Image Pretraining (CLIP) [21]. This approach enables the encoders of a larger predictive system to learn a joint vector space that aligns visual and text-based representations attending to their shared semantics. Using Transformer-based architectures as the spinal cord of these encoders, and training them under a CLIP-like strategy results in feature extractors from which to estimate, for every video and its associated captions, embeddings from both modalities that highlight the common semantics. This data representation has proven useful in several text-image tasks, for example in the case of text-conditioned image generation [7, 22]).

With regards to the prediction media memorability, recent advances have explored which areas of the brain are involved in the process of deciding what content humans most easily remember [10], as well as what are the intrinsic properties of the medium relevant to the process (for instance, exploring the conceptual structure of the input [13] or scene variability [14]). From the computer vision standpoint, efforts have traditionally relied on both low-level visual descriptors [9] and in the extraction of region-wise characteristics from the image employing deep conceptual features. Lately, the emphasis has been put on understanding the connection between the global semantics of an image (its visual constituent elements) and memorability. It has been shown that there exists a close correlation between certain topics and average memorability scores [12]. Therefore, even if many factors contribute to the memorability of a given sample, it seems that the main topic of a video (its semantic unit), extracted from text-based sources like captions, may be used as a proxy material to estimate its semantics and tackle the task of predicting memorability.

In light of this, it seems sensible to walk towards the leverage of the aforementioned capacity of encoders trained under a CLIP scene to learn features that align the semantics of text-based and visual content [11]. However, we seek to learn a representation tailored to the particularities of the problem of predicting media memorability. We believe that by undergoing a CLIP-like fine-tuning, encoders can better estimate features adapted to the idiosyncrasy of memorability-related tasks, hence enabling us to develop models with higher predictive power in this context.

3 PROPOSAL

Our proposed model can be split up into two steps, namely *pre-training* and *regression*, as can be seen in Figure 1. In the pre-training step, vision and text encoders are jointly trained under a CLIP scheme. This way, a visual and a text-based encoder are trained together in order to maximize the cosine similarity of the feature embeddings computed separately for each modality, hence learning to align the representation computed for every image-text pair in a dataset. The Transformer-based models employed in the original CLIP paper serve as our baseline, given they stand as models already trained on a large-scale, generalist dataset to correctly identify image-text pairs [21]. We propose to adapt the representations that are obtained using these encoders to the memorability prediction problem by fine-tuning them on image-text pairs extracted from

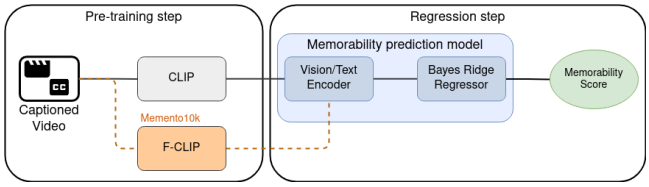


Figure 1: Our pipeline includes a pre-training stage in which baseline encoders are fine-tuned on the memorability dataset following a CLIP scheme, and a regression stage, in which encoders provide a feature representation independent for each modality that is passed to a Bayesian Ridge Regression model in charge of predicting memorability.

videos of a memorability-targeted dataset, namely Memento10K [17]. The resulting fine-tuned CLIP (F-CLIP) models obtained exhibit semantic knowledge related to the visual and textual language that is meaningful to the task. In other words, we *pre-train* encoders to indirectly obtain vector representations useful to tackle the prediction of video memorability from the understanding of its semantics, seen at the same time from both frames and text descriptions.

Once this pre-training step is completed, we decouple these encoders while freezing their learned weights, so they can be used as feature extractors. Their outputs are then fed into a Bayesian Ridge Regressor (BRR), which is the only part of the pipeline that actually predicts memorability scores. Bayesian Ridge Regression is a statistical model that combines the principles of Bayesian inference with Ridge regression, assuming a prior distribution over the regression coefficients and incorporating regularization to mitigate overfitting. It yields similar results as Ordinary Least Squares Linear Regression whilst being more robust to ill-posed problems. The default implementation offered by the *sklearn* library is chosen [20]. Making use of this simple yet robust model allows us to put focus on the comparison between the different extracted features.

4 EXPERIMENTAL SETUP

4.1 The Memento10k Dataset

Memento10K [17] is a dataset of 10,000 short videos which focuses on visual memory and was devised to model video memory decay as a function of the interval between repetitions of a video during a visual memory game. The videos were collected by scraping the Internet, keeping only those regarded by the annotators as "home videos". This incurs a relatively poorer overall image quality. However, the content included in Memento10K places emphasis on human actions and motion, resulting in sudden changes in images and a higher degree of optical flow. Videos are short (around 3 seconds long on average), and thus represent a single semantic unit. The authors observed that the success rate of the participants recognizing a video decayed linearly as a function of the number of projected videos between two repetitions. To account for this linear decay, the raw memorability score of every video (the percentage of people successfully recalling it in a subsequent viewing) was post-processed, hence obtaining memorability labels as the likelihood of a given video being remembered after an interval of 80 clips.

The Memento10K dataset includes Closed Captions (CC), a set of objective summaries written by human annotators. These natural language-based texts barely present any emotional bias, and therefore condense the semantics of the video in a concise manner. There are 5 captions associated with every video, and while they convey similar information, each expresses it differently, making for a rich source of knowledge about the semantics of a scene while enabling a straight-forward association with its purely visual features.

4.2 Two-stage predictive pipeline

In order to test our proposed models on the memorability task, two different approaches were adopted. Firstly, we aimed to predict the memorability score of individual frames or captions of a video, which we refer to as the *frame/caption-level* task. In this task, the label of the entire video was assigned to each of its extracted frames and for every caption describing it. This is akin to an image memorability problem, in which the input comprises still images without any motion or temporal information. Secondly, we tried to predict the memorability score assigned to the entire video (*video-level* task). However, since the input to the visual or textual encoders remained a single image or caption, we compared two pooling methods that produced a single predicted score for the entire video: early and late average. The former computes the mean embedding of all inputs (visual - first, middle and last frames, or textual - the five captions in the set) belonging to a given video, using it to obtain a single output from the BRR. The latter computes a score prediction for every input of a given video and averages the scores to obtain a final prediction. In the following paragraphs we specify the details of each of the steps that comprise the model.

Regarding the pre-training step, the 2021 MediaEval Workshop partition of the dataset was used for experimentation, leaving a set of 7,000 videos for training and 1,500 videos for validation. The remaining test 1,500 videos were not used for training as their labels were not available, resulting in a final corpus size of 8500 videos. We extract the first, middle and last frames of each video (roughly equivalent to 1FPS subsampling), obtaining a total of 25,500 images. Each frame is paired with each of the 5 captions that describe the video, resulting in a total of 127,500 image-caption pairs. Both frames and captions are preprocessed using the default CLIP processor. We use the AdamW optimizer with a learning rate of $1e^{-4}$ and a batch size of 32 samples. An early-stopping criterion was used to terminate training after 5 epochs without improvement.

In the regression step, the weights of the encoders are frozen so they serve as feature extractors. We train and evaluate the BRR following a 5-Fold Cross Validation scheme, reporting on global performance as the mean SRCC for all the folds. Folds are consistent with videos, hence ensuring that all the material associated with a given video (both descriptions and visual frames) is kept together within the same fold.

5 RESULTS AND DISCUSSION

5.1 Pre-training step

As explained in section 3, a CLIP-like strategy was used to train encoders on visual and textual inputs, resulting in an aligned embedding space for both modalities. To assess this alignment after fine-tuning on Memento10k, we extracted embeddings for each

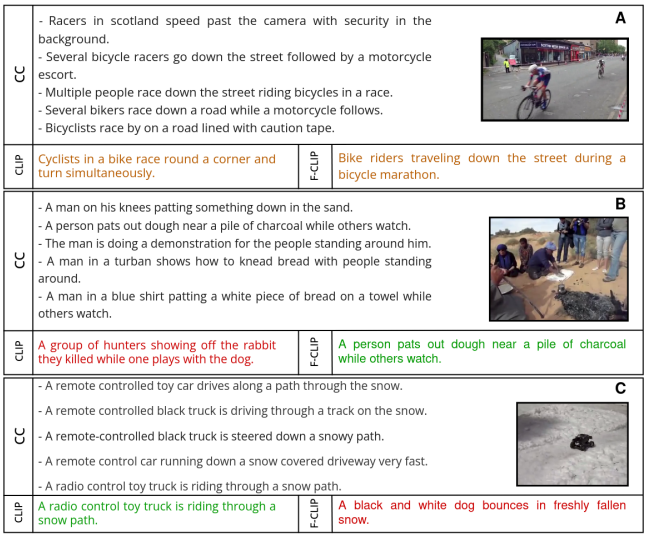
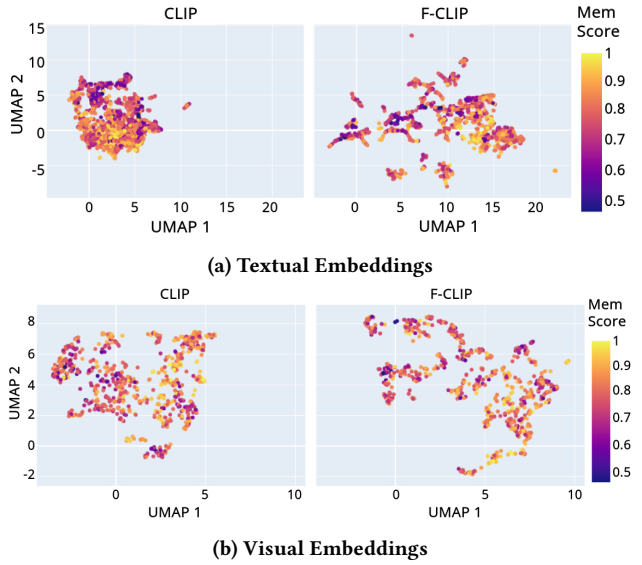


Figure 2: CLIP models learn general descriptions of images and their accompanying text. Green text signifies correct predictions from the CC set, while orange indicates conceptually accurate descriptions that lay outside the CC set. Red captions are neither in the CC nor conceptually accurate.

image and caption in the original test set of the corpus, which can be used here for evaluation as no memorability labels are needed. We computed the fraction of times that an image vector, extracted from a frame of a video, had its closest textual vector (based on cosine distance) belonging to a caption in the same video, resulting in a comparable performance between the baseline and fine-tuned models (49.7% ± 1.5% and 49.4% ± 1.5% respectively).

Figure 2 displays examples considered by this evaluation strategy as errors but providing valuable insights. In Example A the selected caption was not part of the original set, however, the text accurately described the elements and actions in the image, belonging to a similar-themed video in the dataset. In example B, F-CLIP correctly selected the caption while CLIP failed due to image characteristics. The defining element (e.g., "bread" or "dough") was barely visible, limiting visual information. Nevertheless, scene composition and general elements were accurately captured. Example C showed CLIP selecting the correct caption, while F-CLIP accurately captured image colors ("black" and "white"), but struggled with objects of similar shape (e.g., "dog" and "toy car"). These examples suggest that both CLIP and its fine-tuned version can distinguish complex visual elements and understand scene semantics.

We study the effect of transforming the embedding space by projecting the 512-dimensional visual and text-based embeddings into a plane using UMAP [16]. This method is commonly used for dimensionality reduction as it performs a non-linear transformation that keeps most of the spatial information of the original space using manageable computing resources and time. A qualitative analysis shows that, for the text-based branch, this process may help in separating them into roughly heterogeneous regions sharing a similar range of memorability scores (Fig. 3a). The same is not so evident for visual embeddings (Fig. 3b), still, visual inspection



(a) Textual Embeddings

(b) Visual Embeddings

Figure 3: Comparison of CLIP and F-CLIP embeddings projected using UMAP. (a) Fine-tuned text-based embeddings show greater dispersion and exhibit separations between clusters associated with memorability (e.g., dark-colored clusters in upper-left). (b) Visual embeddings display less evident transformation of space, making it harder to identify memorability-related clusters.

leads us to think that there are differences between high and low memorability regions that a linear method can learn. To assess it, we evaluate the fitness of the computed text-based and visual embeddings to enhance the predictive capacities of a relatively simple BRR in the specific scenario of memorability.

5.2 Regression results

In table 1 we compare the results of our F-CLIP embeddings with the representation obtained by the original default CLIP as loaded from the checkpoint (baseline encoders) [1]. Although as a rule, greater SRCC ratios can be noticed when we employ F-CLIP, whereas the text-based branch does experience a significant difference, the same does not hold in the case of visual data. A possible explanation might come from the fact that text is potentially easier to understand from the point of view of a CLIP pre-training scheme. As a matter of fact, we are matching images and their descriptions, but sentences, even if they consist of mere objective descriptions, exhibit more variability than individual frames. This can be understood as text being a more accessible source of information about a sample. Opposed to it, visual information, particularly in Memento10K, whose videos generally display low image quality, offers little variance and hence visual branch struggles to find relevant patterns for that modality.

Experiments on a video-level demonstrate a similar trend of improvement when using F-CLIP, yet not as evident as in the frame/caption-level case. This shows that the benefit of fine-tuning the encoders is lost when pooling is performed, meaning that the decision to average either multiple representations or predicted scores can result in a loss of detail that degrades performance. Moreover, there is no significant performance pattern to choose between pooling methods.

[1]Downloaded from https://huggingface.co/openai/clip-vit-base-patch32

Table 1: SRCC using as feature extractors either baseline CLIP encoders or those fine-tuned by our F-CLIP proposal for each modality. The boldface indicates a significant improvement upon baseline.

Task	Pooling	Model	Text	Vision
Frame/Caption-level	-	CLIP	0.538 ±0.007	0.639 ±0.008
		F-CLIP	**0.575** ±0.007	0.648 ±0.008
Video-level	Early AVG	CLIP	0.615 ±0.021	0.671 ±0.021
		F-CLIP	0.635 ±0.021	0.676 ±0.021
	Late AVG	CLIP	0.602 ±0.021	0.671 ±0.021
		F-CLIP	0.626 ±0.021	0.671 ±0.021

6 CONCLUSIONS AND FUTURE WORK

In this work, we have proposed and validated the use of a fine-tuning step of text-based and visual encoders using CLIP as an effective method for extracting informative features from videos and their textual descriptions to predict video memorability, which we have called fine-tuned CLIP (F-CLIP). CLIP allows the encoders to learn an aligned representation of the semantics contained in images and texts, thus enabling them to perform well in subsequent tasks with little or no further training. Nonetheless, our study has also shown that an additional cross-modal fine-tuning step has an important impact on the prediction of video memorability, leading to more accurate estimations compared to using the general, zero-shot model. These findings suggest that CLIP-like schemes specifically tailored to the target task can be a valuable tool for improving our understanding of what makes certain videos more memorable than others.

However, probably our biggest limitation comes from using individual frames as input omits motion aspects, losing related semantic information. This might be addressed by incorporating motion descriptors [19] or Video Transformers that instead work directly with image sequences [1, 2]. Furthermore, text-based encoders would benefit from some of the capabilities shown by Large Language Models (LLMs) such as PaLM, GPT-4 or LLaMA [4, 18, 23].

Future research would need to explore the ability of contrastive learning methods to represent memorable media, perhaps including explicit information about the task during the pre-training phase. Also, incorporating low-level, pixel-driven statistical descriptors, and mapping them to high-level, semantic information (which we know correlates with memorability) would help us to build a bottom-up knowledge of the phenomenon and broaden the range of applications of this technology.

ACKNOWLEDGMENTS

The work leading to these results was supported by the European Commission through the project ASTOUND (101071191—HORIZON-EIC-2021-PATHFINDERCHALLENGES-01). I. Martín-Fernández's research was supported by the UPM (Programa Propio I+D+i).

REFERENCES

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. arXiv:2103.15691 [cs.CV]

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding? arXiv:2102.05095 [cs.CV]

[3] Zoya Bylinskii, Lore Goetschalckx, Anelise Newman, and Aude Oliva. 2022. Memorability: An image-computable measure of information utility. *Human Perception of Visual Information: Psychological and Computational Perspectives* (2022), 207–239.

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311 [cs.CL]

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]

[7] Federico Galatolo, Mario Cimino, and Gigliola Vaglini. 2021. Generating Images from Caption and Vice Versa via CLIP-Guided Generative Latent Space Search. In *Proceedings of the International Conference on Image Processing and Vision Engineering*. SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0010503701660174

[8] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. 2011. Understanding the intrinsic memorability of images. *Advances in neural information processing systems* 24 (2011).

[9] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable?. In *CVPR 2011*. IEEE, 145–152.

[10] Andrew Jaegle, Vahid Mehrpour, Yalda Mohsenzadeh, Travis Meyer, Aude Oliva, and Nicole Rust. 2019. Population response magnitude variation in inferotemporal cortex predicts image memorability. *Elife* 8 (2019), e47596.

[11] Ricardo Kleinlein, Cristina Luna-Jiménez, and Fernando Fernández-Martínez. 2021. THAU-UPM at MediaEval 2021: From Video Semantics To Memorability Using Pretrained Transformers. In *MediaEval Multimedia Benchmark Workshop Working Notes*.

[12] Ricardo Kleinlein, Cristina Luna-Jiménez, David Arias-Cuadrado, Javier Ferreiros, and Fernando Fernández-Martínez. 2021. Topic-Oriented Text Features Can Match Visual Deep Models of Video Memorability. *Applied Sciences* 11, 16 (Aug 2021), 7406. https://doi.org/10.3390/app11167406

[13] Talia Konkle, Timothy F Brady, George A Alvarez, and Aude Oliva. 2010. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of experimental Psychology: general* 139, 3 (2010), 558.

[14] Talia Konkle, Timothy F Brady, George A Alvarez, and Aude Oliva. 2010. Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological science* 21, 11 (2010), 1551–1556.

[15] Qi Lin, Sami R Yousif, Marvin M Chun, and Brian J Scholl. 2021. Visual memorability in the absence of semantic content. *Cognition* 212 (2021), 104714.

[16] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[17] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. 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*. Springer, 223–240.

[18] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[19] Reinier Oves García, Eduardo F Morales, and L Enrique Sucar. 2021. Second-order motion descriptors for efficient action recognition. *Pattern Analysis and Applications* 24 (2021), 473–482.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[25] Weizhen Xie, Wilma A Bainbridge, Sara K Inati, Chris I Baker, and Kareem A Zaghloul. 2020. Memorability of words in arbitrary verbal associations modulates memory retrieval in the anterior temporal lobe. *Nature human behaviour* 4, 9 (2020), 937–948.