Long-Term Ad Memorability: Understanding & Generating Memorable Ads

Harini SI* **Somesh Singh*** **Yaman K Singla*** **Aanisha Bhattacharyya**
Veeky Baths **Changyou Chen** **Rajiv Ratn Shah** **Balaji Krishnamurthy**
Adobe Media and Data Science Research, State University of New York at Buffalo, IIIT-Delhi, BITS Pilani, Goa

Abstract

Despite the importance of long-term memory in marketing and brand building, until now, there has been no large-scale study on the memorability of ads. All previous memorability studies have been conducted on short-term recall on specific content types like action videos. On the other hand, long-term memorability is crucial for advertising industry, and ads are almost always highly multimodal. Therefore, we release the first memorability dataset, LAMBDA, consisting of 1749 participants and 2205 ads covering 276 brands. Running statistical tests over different participant subpopulations and ad types, we find many interesting insights into what makes an ad memorable, e.g., fast-moving ads are more memorable than those with slower scenes; people who use ad-blockers remember a lower number of ads than those who don't. Next, we present a model, Henry, to predict the memorability of a content. Henry achieves state-of-the-art performance across all prominent literature memorability datasets. It shows strong generalization performance with better results in 0-shot on unseen datasets. Finally, with the intent of memorable ad generation, we present a scalable method to build a high-quality memorable ad generation model by leveraging automatically annotated data. Our approach, SEED (Self rEwarding mEmorability Modeling), starts with a language model trained on LAMBDA as seed data and progressively trains an LLM to generate more memorable ads. We show that the generated advertisements have 44% higher memorability scores than the original ads. We release this large-scale ad dataset, UltraLAMBDA, consisting of 5 million ads. Our code and the datasets, LAMBDA and UltraLAMBDA, are open-sourced at https://behavior-in-the-wild.github.io/memorability.

1. Introduction

"The first lesson of branding: memorability. It is very difficult buying something you can't remember." - Sir John Hegarty, the creator of the iconic ads for Levi's, Nike, Microsoft, Tinder, and Coke.

The global advertising industry is $700 billion+ industry [23]. Three out of the ten largest companies by market capitalization are advertising companies with average revenues exceeding $250 billion. The World Wide Web is mostly funded by advertising. Given that marketers are spending such large sums of money on advertisements, it is imperative to know if their brand would even be recalled at the customer's purchase time. This would help the marketers optimize their costs, content, delivery, and audience, ultimately helping in boosting sales. Most of the studies carried out in the machine learning literature have been on short-term memorability (memorability testing in less than 5 minutes) on action videos like walking and dancing (Table 1). On the other hand, customer purchase decisions are rarely carried out within five minutes of watching an ad. In fact, the marketing funnel model popular in the marketing literature says that customers pass through several stages of a funnel, like awareness and consideration, before the actual sale [42]. Further, in the ML literature, there have been no memorability studies on advertisements. Advertisements are highly multimodal; they contain video, speech, music, text overlaid on scenes, jingles, specific brand colors, *etc*. None of these elements are found in previous studies like VideoMem, Memento10k, LaMem, *etc.* (refer to Table 1 for a detailed comparison).

What drives memory? Memory over content is determined by two factors: human factors and the content itself [12]. Human factors represent the viewer's thoughts, emotions, and actions, while the content factors are words and raw pixels of text, images, and videos. Foundational large-scale studies on memorability [2, 16, 30, 36] showed that there is sufficient consistency between humans in what they remember. Human-human memorability consistency scores are in the range of 0.6-0.8. This means that the memorability ranks of a content between two groups of humans are more than 60% correlated.

These initial studies also tried to answer the question of what makes a content memorable. They found that low-level image features like colors, aesthetics, number of objects, and such have very little correlation with whether the image

*Equal Contribution. Contact behavior-in-the-wild@googlegroups.com for questions and suggestions

Dataset	#Samples	Memory Type	Memory Retrieval Process	Content	Average Screen Duration	Audio Present	Human Consistency	Memorability Measurement Protocol
Memento10k	10,000	ST ($<$ 10 mins)	Recognition	Videos of single type of action obtained from amateur videos	3s	Yes	0.73	Competition
VideoMem	10,000	ST (few mins), LT (1-3 days)	Recognition	Videos of a single type of action obtained from professional (staged) footage	7s	None	0.48 (ST), 0.19 (LT)	Competition
LaMem	60,000	ST ($<$ 3.5 mins)	Recognition	General Images	0.6s	None	0.68	Competition
SUN	2,222	ST ($<$ 4.4 mins)	Recognition	General Images	1s	None	0.75	Competition
MemCat	10,000	ST ($<$ 3.5 mins)	Recognition	General Images	0.6s	None	0.78	Competition
MediaEval	1500	ST (few mins) and LT ($<$ 3 days)	Recognition	Short video clips collected from Twitter and Flickr	6s	None	-	Competition
LAMBDA (Ours)	2,205	LT (1-3 days)	**Recognition and Recall**	Videos of multimodal advertisements	**33s**	**Yes**	**0.61**	**Natural**

Table 1. Comparison of all the major (image and video) memorability datasets available in the literature along with LAMBDA (ours). The datasets are compared on the following axes: number of samples, type of memorability (short-term (ST) and long-term (LT)), memory retrieval process (recall or recognition), type of content (images/videos and their type), duration with which the sample was shown on the participants' screen, whether audio was present or not, human consistency achieved in the study, and the protocol followed in the study to collect the data. **Memento10k** - [51], **VideoMem** - [16], **LaMem** - [36], **SUN** - [30], **MemCat** - [27], **MediaEval** - [38]

was remembered. On the other hand, high-level features like object and scene semantics have significant correlation with memorability. For example, human images are more memorable than object images. Further, these initial studies contributed to protocols for conducting memorability studies. They proposed a competitive memorability game, where they asked participants to recognize as many images as they could remember. The game ended for those participants whose scores fell below certain success rate thresholds. However, this protocol limits the scope of these studies to short-term memorability (a few seconds to a few minutes), and the competitive nature makes the study unnatural and, thus, not applicable to real-world scenarios like marketing where the customers are not competing with each other to remember the brand. Therefore participants in all these studies are aware that they are being tested for memorability, this can create a deviation from their natural behaviour commonly known as the Hawthorne effect in psychology [49, 50, 61]

What drives customer memory? Customer purchase decision is a long process. Marketing theory formulates this as a funnel where customers pass through several stages like awareness, consideration, and evaluation before the actual sale [42]. Due to the purchase funnel being a multi-stage lengthy process, long-term memorability (LTM) is the closest proxy to model customer memory [53, 71]. While the LTM store (as distinct from the STM store) has been studied for over five decades in psychology [5, 21], there are no large-scale studies containing data over such time period that can help us model the long-term customer LTM spanning days or more [53, 71]. Unfortunately, STM datasets, typically measuring memorability of a few seconds to a few minutes, are not good proxies to model customer memory. Moreover, the competitive nature of the memorability games in the previous studies further disconnect the modeling from advertising use cases.

To answer the question of what drives customer memory, there have been efforts in marketing literature where researchers have conducted many field experiments with the intent to prove certain hypotheses. For instance, Li *et al.* [43] conducted a field experiment on advertisements shown during the 2006 Super Bowl Games where they asked the audience to recall the brands they saw in the game held (at least) a day earlier. They found a strong primacy effect, where viewers remembered brands more if they occurred earlier when controlling for the commercial length. Similarly, there have been studies to determine the effect of syntactic complexity [4], emotional content [48, 55], repetition [62], spot length [52, 70], the position of brand logo and imagery [52], sound presence [7], and on customer factors like involvement and relevance [52, 62].

While these studies have contributed much towards understanding the factors that drive customer memory, they are limited in their scope. These field experiments evaluate the effect of a single content factor while controlling for others. Further, these are conducted on a small number of advertisements.Therefore, to model LTM over advertisements, we conduct the first large-scale human study on long-term advertisement memorability[1]. We call it LAMBDA (Long-term Ad MemoraBility DAtaset). Over two years, we conducted an LTM study involving 1749 participants across four sessions across two institutes to collect LAMBDA. We collect memorability scores over 2205 ads from 276 brands, covering 113 industries. On day 1, participants saw ads, and after a lag time of at least one day, they answered questions testing their brand recall, ad recall and recognition, scene recall and recognition, and audio recall (§2.2). Next, we average the brand recall scores across participants and compute the average long-term ad memorability scores. Then, we use these scores to train machine learning models to predict long-term ad memorability.

How can we model customer memory? To model customer memory, we design a novel architecture, Henry[2]

[1]We obtained the Institutional Review Board Approval to conduct the study from our institute.

[2]We name the model Henry in honor of the immense contributions by the patient Henry Molaison (H.M.) [67]. An experimental surgery conducted on him resulted in the discovery of the distinct regions responsible for LTM

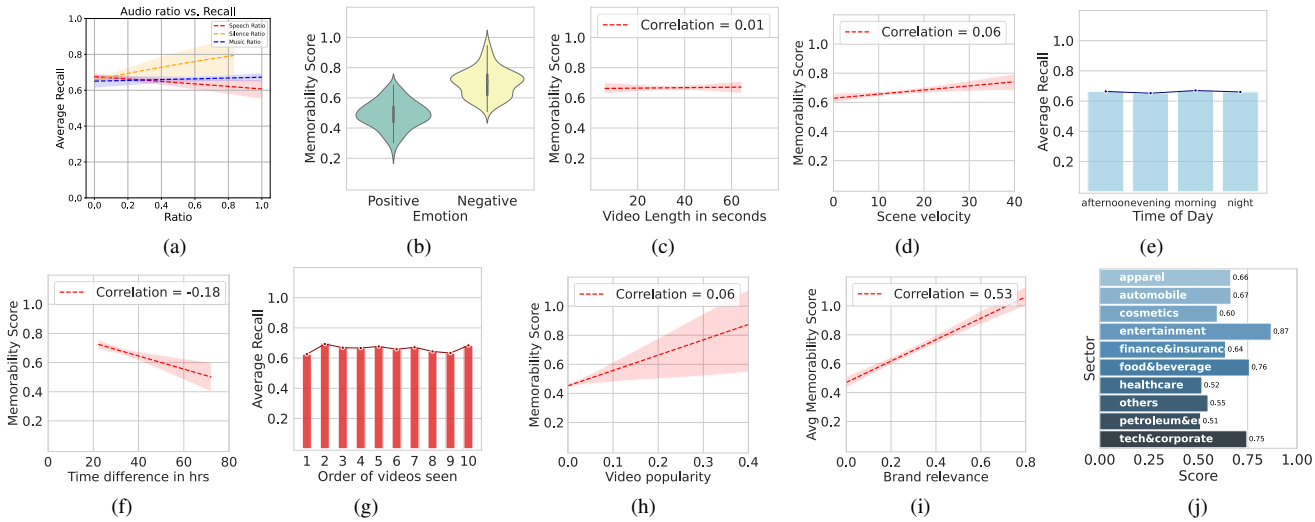

Figure 1. Correlations between *content factors* (a-d), *interaction factors* (e-g), and *customer behavior factors* (h-j) with memorability on LAMBDA samples. While emotion has a high correlation with memory, other content factors do not have much correlation. Further, while there is little correlation between the order of videos seen and memorability; with time, participants' memory of the videos shows a forgetting trend. Video popularity, as measured by YouTube likes/views, shows a slight positive correlation with memory. Average brand relevance has a strong positive correlation with memory, with top sectors being remembered as food, entertainment, and tech. Speech, silence and music have little effect with silence having the highest positive correlation with recall. Silence ratio is measured as the percentage of silence in a video, similarly for music and speech.

(Fig. 2), incorporating world-knowledge from large language models (Llama [69]), visual knowledge from vision encoder (EVA-CLIP [68]) and specialized perception modules covering visual and cognitive knowledge about the ad. The world knowledge helps Henry to understand the semantics of the ad, the brand knowledge and consolidate them with the visual semantics from the ad. The visual encoder helps the model to "see" the ad. We convert the visual encoder embeddings to language space using QFormer [44] and further augment them with specialized "verbalizations" involving visual scene descriptors like visual caption, optical character recognition (OCR), automatic speech recognition (ASR), and cognitive descriptors like emotion and scene complexity scores, which help the model ground the visual and cognitive knowledge in the LLM's world knowledge. We train the model on our LTM data samples and obtain higher than human consistency scores. Further, we train Henry on other short and long term image and video memorability datasets in the literature - LaMem, MemCat, SUN, Memento10k, MediaEval, and obtain state-of-the-art performance on all of them. We also show that Henry performs well on unseen datasets in zero-shot settings, performing better than models specifically trained on those datasets.

How to generate memorable Ads? One of the primary goals of modeling content memorability is to generate more memorable content. The task of generating more memorable ads is given the ad description containing the brand and campaign title to generate the ad scenes and dialogues. However, there is no data in the literature for this task. Therefore, we turn to synthetic data generation and LLM-as-a-judge paradigm [34, 75]. We first collect a large-scale advertisements dataset, collecting brand name, ad text, time, ad content, and channel. Then, we use Henry as a judge to simulate memorability on the collected ads. We ultimately get a dataset of 5 million advertisements with their automatic speech transcripts, OCR, automatically detected objects, colors, aesthetics, captions, emotions, logos, and memorability scores. We call this dataset UltraLAMBDA. We then select high memorability samples from UltraLAMBDA to train Llama-13B to generate memorable ads. Finetuning LLama for two iterations on this automatically constructed dataset yields an improvement of 44% in memorable ad generation.

Our main contributions are summarized as follows:

- We release the first large-scale dataset, LAMBDA, on long-term advertisement memorability involving more than 1700 participants. We collect memorability scores over 2205 ads from 276 brands (157/276 brands are from SnP 500), covering 113 industries. Further, we introduce a new protocol to measure customer memory of brands (§2.2).

- We design a novel model, Henry, which can model both STM and LTM and can incorporate scene understanding, brand knowledge, and speech (§3). Henry achieves state-of-the-art performance on eight literature image and video memorability datasets (§3.3). Further, we show that Henry performs well on unseen datasets in zero-shot settings.

and STM.

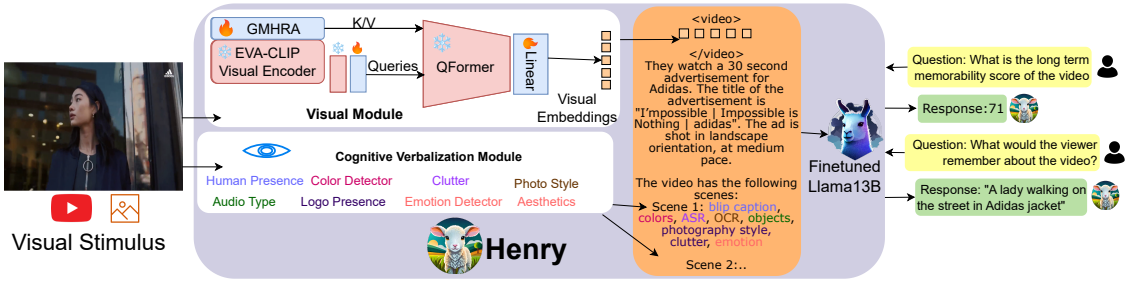

Figure 2. Predicting memorability by encoding visual information (via visual encoder EVA-CLIP), cognitive concepts (via verbalization module), and world knowledge (through fine-tuned Llama). We instruction fine-tune the combined model end to end to predict user memorability. Snowflake and fire symbols denote the frozen and unfrozen parts of the architecture.

- We propose the task of memorable ad generation. We release the first large scale ad dataset, UltraLAMBDA, consisting of 5 million ads with their automatically extracted content labels like ASR, captions, OCR, emotions, and memorability scores assigned by Henry. Using UltraLAMBDA, we first show that large LLMs like GPT-3.5 and 4 are unable to generate memorable content. Then, we train Henry to progressively generate more memorable ads resulting an average improvement of 44% in memorability scores (§4). Through this, for the first time in literature, we also show the use of synthetic data on a task for which no large scale data exists.
- We conduct an extensive set of experiments on memorability prediction, showing the effects of LTM on STM modeling and vice-versa, and the effects of changing world-knowledge with time, scene understanding, brand knowledge, and speech on memorability modeling (§3.3).

2. LAMBDA Protocol, Study & Insights

We first give an overview of LAMBDA data collection process and the annotation protocol. We also present some interesting characteristics LAMBDA exhibits about LTM.

2.1. Video Collection

In contrast to previous video memorability works where videos were soundless and only of action videos [16, 51], the videos in our dataset come from multimodal ads released on YouTube channels of 276 major brands covering 113 industries. We collect 2205 such ads spanning over the years 2008-2023. The videos have an average duration of 33 seconds. Out of all the videos, 2175 have audio in them. The collected advertisement videos have a variety of characteristics, including different scene velocities, human presence and animations, visual and audio branding, a variety of emotions, scene complexity, and audio types.

2.2. Annotation Protocol

At the outset, participants are given a preliminary questionnaire aimed at establishing their brand-related interactions and media consumption habits. Participants are given a list of fifteen randomly chosen brand options and are asked to choose those they recall encountering advertisements for during the current year. Subsequently, participants are presented with another set of fifteen brands and are instructed to identify those for which they have personally utilized products within the same timeframe.

In addition, participants are asked about their utilization of ad-blocking software and their Youtube subscription. The questionnaire further captures participants' digital media habits, including the division of their time spent on YouTube between mobile and web platforms and their preferred channels for acquiring information about new products and brands.

Following the questionnaire, participants proceed to the core segment of the study, where they are shown 11 advertisements sequentially. Notably, the eleventh advertisement is deliberately repeated for half of the participants, while it is unique for the other half. To ensure participant engagement, attention-check questions are placed between every two to three advertisements. These questions are common sense questions like "How many legs does a cow have?". If the participant fails to answer the question within 10 secs, they are requested to rewatch the video. After the 11th video, participants are asked if they recollect watching the ad in the span of the study. Interestingly, 15% participants were not able to recognize the repeated video correctly.

The memorability test included 1,749 participants: 971 in a take-home setting and 778 in an auditorium. Take-home participants received questionnaires via email 24 hours after exposure, with responses accepted within a 72-hour window. Auditorium participants completed questionnaires at intervals of 24, 36, or 72 hours, with equal distribution across these time points. The questionnaire assessed two types of memory: brand recognition and ad recall. For recognition, participants identified previously encountered brands from a randomized list of 20. For recall, they described the advertisements associated with the brands they recognized in the earlier prompt[3].

The average memorability score was 67.5% (SD=13.6%).

[3]The complete questionnaire for participant one is given in Appendix:§F.1.

Models	Image Datasets				Video Datasets			
	Lamem	Memcat	SUN	Merged	Memento10k	VideoMem	MediaEval	LAMBDA
Human Consistency	0.68	0.78	0.75	-	0.73	0.61	-	0.55
10-shot GPT-3.5	0.29	0.18	0.15	-	0.07	0.06	0.06	0.06
Regression using ViT feats (ViTMem)	0.71	0.65	0.63	0.77	0.56	0.51	-	0.08
Current Literature SOTA	0.71	0.65	0.68	0.77	0.67	0.56	0.46	-
Henry trained on individual datasets	0.74	0.82	0.73	-	0.75	0.64	0.50	0.55
Henry trained on all (combined) datasets	0.72	0.79	0.76	0.79	0.72	0.60	0.48	0.52

Table 2. Results of Henry (our model) on eight datasets compared with the current best models reported in the literature and GPT-3.5. Human consistency values are also listed in the top row for reference. It can be observed that our model achieves state-of-the-art performance across all datasets. Best models are denoted in green and runner-ups in blue. References for the seven literature SOTA models in the format {`dataset: SOTA model citation`} are: LaMem: [28], MemCat: [28], SUN: [22], Merged Image datasets: [28], Memento10k: [20], VideoMem: [20], MediaEval: [47]

To assess human consistency, we divided the participant pool into two independent halves and measured the agreement between memorability scores from one half and the other. Across 25 random split-half trials, we obtained a Spearman's rank correlation (ρ) of 0.77 for brand recall (compared to $\rho = 0.68$ for images [36], $\rho = 0.616$ for videos in [16], and $\rho = 0.73$ in [51]). The estimated sensitivity index (d') for participants was calculated as 1.848 [72].

2.3. What makes an Ad memorable?

Among the many reasons why an ad might be memorable, we investigate the following factors: **brand factors** (*viz.,* brand popularity, industry), **content factors** (*viz.,* video emotion, scene velocity, length, speech to silence ratio), **customer-content interaction factors** (*viz.,* time of seeing the video, order in which the video was seen, time difference between watching the video and recalling the brand), and **customer behavior factors** (*viz.,* average relevance of the brand and video popularity).

Content Factors: Previous studies like [30, 51] have investigated the effect of pixel statistics like color and hue, saturation, and value, scene semantics like the number of objects, the area occupied by objects on memorability. In general, low-level semantic features have no correlation with memorability, but higher-level features like the type of scenery has some correlation. For instance, Newman *et al.* [51] found that videos with people, faces, hands, man-made spaces, and moving objects are, in general, more memorable than those with outdoor landscapes or dark and cluttered content. Since only our dataset has videos with cognitive features like emotions and are also non-silent, we extend the previous analysis to find the effect of speech and emotion on memory. Fig. 1a shows the effect of speech. We observe that percentage of speech in a video, presence of music, and type of music have a very little correlation with long term memory. On the other hand, emotions primarily depicted through speech in ads can explain memorability. We see in Fig. 1b that negative emotions are more memorable than positive emotions. Further, we find that video length has little effect on memorability (Fig. 1c), but scene velocity has a slightly positive correlation with memory (Fig. 1d).

Interaction Factors: Memorability may also depend on the time of the day the ad was seen. However, we find that the time of day of watching has almost no effect on the memorability of the ad (Fig. 1e). It may be expected that memorability decays as time passes. We plot the forgetting curve for ads in Fig. 1f measuring brand recognition against time elapsed between video viewing and memory assessment. The forgetting coefficient of ads is 0.18, notably than action videos [16]. The difference likely arises due to differences in protocols. Cohendet *et al.* (2019) [16] used a two-stage memory protocol in which participants did both short-term and long-term recall, thus enhancing their long-term recall. Next, we investigate the effect of the order in which the video was watched with its memorability (Fig. 1g). We see that order of videos seen has little impact on video memorability, with a slight bias in favor of the initial and last ads.

Customer Behavior Factors: It might be possible that the videos which are liked more are remembered more. To investigate this, we test the correlation of popularity as measured by the ratio of Youtube video likes to views with memorability. We see that there is a positive correlation between video popularity and memorability (Fig. 1h). Further, in the study, we asked the participants to select the brands they have personally used from a set of 15 randomly chosen brands and similarly choose brands they have seen ads for. To prevent any systematic bias, the brands asked in this question are independent of the brands shown the next day. We plot thus collected brand relevance values with brand recall in Fig. 1i. We see that average brand relevance is strongly correlated with average recall (coeff= 0.53), where entertainment, corporate, and food and beverage sectors, which are quite popular brands in a student population are the most remembered, while the others are less remembered (Fig. 1j).

3. Predicting Ad Memorability

In this section, we focus on predicting memorability - both long-term and short-term for both videos and images. We pose memorability prediction as a problem that needs (a) *visual knowledge* to identify and understand visual con-

cepts across images and videos like shapes, colors, objects, and scenes, (b) *cognitive knowledge* relevant to marketing, for example, ad emotions, scene complexity, scene aesthetics, and (c) *world knowledge* to relate the captured visual and marketing concepts to real-world concepts capturing their function, use, and interaction patterns. For instance, when Airbnb[4] shows an adult female and a male with the text, "Our guest room is paying for our wedding"; it denotes a couple saying that renting out their space on Airbnb helps them sponsor their wedding [40]. World knowledge captured in LLMs, together with the visual knowledge of ViT and marketing knowledge through specialized cognitive models, helps to (i) identify the two adults as a couple, (ii) AirBnb as a housing company, (iii) recognize the warm emotional tone of the text, and make sense of all three concepts together. Fig. 2 shows the proposed architecture of Henry.

3.1. Encoding Multimodal Content

The primary goal of this step is to effectively leverage the "world-knowledge" capabilities of the pre-trained LLM. We choose Llama [69] as our base LLM. We employ two techniques to convert visual data into language: encoding visual frames into the LLM space and verbalizing cognitive concepts into language space. We detail the two steps next.

Sampling Frames: We detect scene changes by analyzing changes in HSV intensity and edges in the scene, with a 0.3 threshold. We choose the threshold value from the 30-degree rule inspired by the concept of jump-cut avoidance in cinematography [3, 24]. The 30-degree rule can be formulated as follows: after a "cut" (camera stops and re-starts shooting), the camera angle must change by at least 30 degrees. For dominant frame selection common blur/sharpness heuristics fail in presence of text in image. So we extract the frame with the least changes using [73].

Encoding Into Language Embedding Space: To give visual knowledge to Henry, we use EVA-CLIP visual embedder [68]. We find that Global Multi-Head Relation Aggregator (GMHRA) [45] helps aggregate the ViT's information better across the time dimension. Next, to effectively leverage the LLM's rich language representations, we use a pretrained Q-Former from BLIP-2 [44] with an extra linear layer and additional query tokens to convert from visual tokens to language tokens.

Verbalizing Cognitive, Experimental, Visual Concepts While visual content encodings are a good representation of the visual characteristics of the image, we find that they are still unable to capture rich cognitive and semantic information present in images. Therefore, to augment the cognitive understanding of the LLM, we verbalize the frame semantic information using the set of features that came out important in our memorability analysis (Fig. 1) [8, 66]. The cognitive and visual features are given in Table 5 and Listing 14. We find that our cognitive verbalization helps ground the visual perception of LLM in the marketing concepts of the image, helping in downstream prediction performance (Appendix:Table 7).

3.2. Two-Stage Training

We do two-stage training where in the first stage, we utilize the Webvid [6], COCO caption [15], Visual Genome [39], CC3M [63], and CC12M [13] datasets to align the visual encoder embeddings with LLM via a large-scale pre-training approach. In the second stage, we train the model with high-quality memorability instructions prepared by following the approach described in the last paragraphs. Henry takes the concatenated inputs, representing the contextual information, and is trained to predict the memorability score of the given image or video within the range of 00 to 99 (see Appendix:Listing 14). The memorability score of a video, is the percentage of times the participants recall the video correctly (we normalise it to an integer value between 00 and 99 to facilitate the LLM training). During training, the LLM predicts from the complete vocabulary, while during inference, we use the softmax function over numeric tokens only to obtain a number.

3.3. Results and Discussion

We conduct extensive experiments on all literature datasets, covering both videos and images, STM and LTM. We compare Henry[5] with the current state-of-the-art models in the literature across eight datasets, including 10-shot GPT-3.5 (text-davinci-003) [54] where we provide GPT with the same verbalization (for 10 examples), as we provided to Henry, as well as with prior regression based methods using features extracted from ViT L-14 [28]. Results are shown in Table 2, which demonstrate that Henry outperforms all the seven models in the literature across all the seven datasets. We also conduct extensive ablations to understand the effect of different kinds of data and architectural choices. They are discussed in Appendix:§C.1.

4. Generating Memorable Ads

We introduce the new task of memorable ad generation. Given inputs like a brand name, a brief campaign description, and the desired ad duration, the goal is to generate a memorable ad featuring scene descriptions, characters, and dialogues. While most memorability research focuses on assessing how memorable content is, little attention has been given to generating memorable content [17, 26, 35, 41, 64]. This gap exists primarily due to the lack of a sufficiently

[4] See Appendix:Fig. 9) for the ad

[5] Computing infrastructure used to conduct the experiments along with hyperparameters are given in Appendix:§H.1. All experiments are conducted with three random seeds and averages are reported.

Model	# Params	Training	Dataset	High Quality Mem Samples	Δ Memorability				Ad-Quality		
					Low	Med	High	Avg	GPT-4 Consistency	GPT-4 Preference	Human-Preference
GPT-4 5-shot	>175B	ICL	$LAMBDA_{High}$	5	+48	+18	-13	+17.6	7.73	91.3%	41.8%
GPT-3.5 5-shot	175B	ICL	$LAMBDA_{High}$	5	+35	+5	-31	+3	7.17	84.2%	-
GPT-3.5 3-shot	175B	ICL	$LAMBDA_{High}$	3	+34	+6	-32	+2.6	6.98	83.1%	-
Henry-SEED	13B	SEED	$UltraLAMBDA$	800k	+41	+18	+1	+20	7.34	74.7%	-
Henry-SEED	13B	SEED	$UltraLAMBDA + LAMBDA_{High}$	820k	+89	+31	+12	+44	7.44	85.6%	60.48%
Henry-SEED	13B	SEED	$LAMBDA_{High}$	650	+78	+13	+1	+30.6	5.03	63.9%	-
Henry-SEED	13B	SEED	$UltraLAMBDA$	50k	+12	+9	-6	+5	6.01	66.1%	-
Henry-SEED	13B	SEED	$UltraLAMBDA$ (w/o high-mem filtering)	2M	+19	+5	-45	-7	6.73	71.1%	-

Table 3. **Ad Generation**: Results of Henry-SEED compared with in-context-learning (ICL) GPT-3.5, 4 on Ad-Memorability and Ad generation quality. See §4 for details of the metrics computed. We see that Henry-SEED generated ads are more memorable than ads generated using 15x larger GPT-3.5 and GPT-4. We test ad quality using GPT-4 as judge and then test the top-two models using human annotators. GPT-4 as a judge rates GPT-4 and Henry-SEED as the top two models. Subsequently, we ask humans to select between the original and generated ad stories. We observed that human annotators preferred Henry-SEED ads more than the original ads 3/5 times, while GPT-4 generated ads are preferred 2/5 times over the original ads. Further, we note that an increase in the amount of training data for Henry-SEED increases its performance across all metrics. Figs. 3-6 and Listings 1-10 contain some qualitative samples generated using Henry-SEED.

large dataset for training models to generate memorable ads. To address this, we release a large-scale dataset of raw ads and propose the Self-rEwarding mEmorability moDeling (SEED) method, which leverages raw ads to create memorable ones.

SEED method (Fig. 4): *Step 1: Self-Instruction Creation:* We gather a dataset of 5 million raw ads sourced from social media platforms, including Facebook, Twitter, Snapchat, and YouTube. For each ad, we collect the brand name, ad title, links, captions, dates, and ad assets (videos and images).

Step 2: Self-Curation: Since these ads are publicly sourced, we employ few-shot Mistral-7B [32] to clean and filter the ads, ensuring they are marketing-focused, semantically relevant, and use proper language (Listing 16). We then automatically label the ads with cognitive features critical for modeling memorability (Table 5). Subsequently, we use Henry to label the ads for memorability scores. This results in a dataset we call *UltraLAMBDA*, from which we select high-memorability ads with scores above 65.

Step 3: Instruction Fine-Tuning: We then train LLaMA-13B to perform two tasks simultaneously: behavior simulation (predicting ad memorability based on ad content; Listing 14) and content simulation (generating ad scenes and dialogues from a brand name, ad title, and required duration; Listing 15). We refer to the model trained using the SEED process as Henry-SEED (Fig. 4).

4.1. Evaluation

We assess the generated ads using four key metrics: (1) memorability, as determined by Henry-Oracle[6], (2) memorability evaluated using perplexity of the generative models on ground-truth high/medium/low ads, (3) ad quality as judged by GPT-4, and (4) ad quality as evaluated by humans. Although content memorability is assessed by average human recall, it is important to note that humans cannot accurately predict how memorable content will be for others [29]. A true test of memorability for generated ads would require a memorability study akin to LAMBDA, which is costly and unscalable due to the number of models and generated ads. Therefore, we measure the memorability of generated ads using two approaches: Henry-Oracle and perplexity on ground truth memorable ads in LAMDA.

In evaluation using *Henry-Oracle*, the expectation is that the generated ad's memorability should be at par with high-memorable samples (score>65) and better than the low (score<44) and medium memorability samples (44<score<65). Perplexity on ground truth low and high memorable ads evaluates the generative model's propensity to generate more memorable content. A stronger model should have a lower perplexity on more memorable content than less memorable content (refer §E for details on perplexity evaluation).

Figure 3. Henry-SEED Prompt: *Generate the detailed description of a 30-second memorable advertisement titled "Brainly Keep Learning 30sec Final 16x9" for the brand Brainly*. Link to the original ad: https://www.youtube.com/watch?v=kytRXyWXivU Original Memorability score: 85. Memorability score of Generated Ad: 99.

[6]The Henry model trained on the complete (test+train sets) LAMBDA.

Using *GPT-4 as judge*, we test two ad-quality metrics: *consistency* and *preference*. Consistency assesses how coherent the generated story is—both internally (e.g., between dialogues) and in relation to the provided brand information and title (Listing 12). Preference measures how often GPT-4 favors the generated story over the original (Listing 11). In *human evaluation*, we ask human annotators to select between the generated and the original ad stories without revealing which is which (§D). This evaluation is conducted with 20 non-expert annotators and 3 ad industry experts with over 5 years of experience in the creative industry. The expectation is that the quality of synthetic ads should be comparable to that of the original ads.

4.2. Results

We compare the following models to generate memorable ads: LLaVA model trained on UltraLAMBDA (we refer to this model as Henry-SEED), GPT-3.5, and GPT-4. GPT-3.5 and 4 are LLMs with strong generative capabilities with high performance across many benchmarks [11].

Evaluation of memorability of the generated ads: Table 3 compares models based on the average increase in memorability, as evaluated by the Oracle model trained on both the train and test sets. Table 4 presents the perplexity of LLaVA before and after training on UltraLAMBDA. Notably, Henry-SEED, trained on UltraLAMBDA, achieves significant improvements in memorability scores across all categories (Low, Medium, and High). In contrast, while GPT-4 and GPT-3.5—despite being 15x larger—enhance the memorability of ads with initially low ratings, they reduce the memorability of ads with high ratings. Table 4 further highlights differences between untrained and SEED-trained LLaVA. The SEED method substantially lowers perplexity on high-memorability samples. While the original LLaVA model exhibited higher perplexity for high-memorability samples, training on UltraLAMBDA reverses this trend: perplexity increases for low-memorability samples and decreases for high-memorability ones. This shift suggests that the SEED approach enhances the generation of high-memorability ads while simultaneously reducing the likelihood of producing low-memorability ones.

Importantly, UltraLAMBDA contains no overlap with LAMBDA. Neither Henry (used to label memorability for UltraLAMBDA) nor Henry-SEED (trained on UltraLAMBDA) was trained on LAMBDA's test-set ads. Despite this, Henry-SEED demonstrates significant improvement in performance compared to GPT-3.5 and GPT-4.

Evaluation of the quality of the generated ads: When comparing ad quality, we find that while GPT-4 favors its own generated ads 91.3% of the time, Henry-SEED follows closely with an 85.6% preference score. In human evaluations, where annotators were asked to choose between original and generated ads based on quality, Henry-SEED's ads were preferred around 60% of the time—approximately 20% more than GPT-4's ads.

Model	Training	Low(↑)	Medium	High(↓)
LLaVA	0-shot	5.08	5.11	5.39
Henry-SEED	LAMBDA$_{HIGH}$	6.07	3.01	2.17
Henry-SEED	UltraLAMBDA	7.09	4.51	2.35

Table 4. **Ad Generation**: Perplexity comparison (refer §E) of LLaVA and Henry-SEED on low/medium/high memorable ads from LAMBDA test set. We see that untrained LLaVA does not favor memorable ads. Further, we note that when synthetic data is included during training, the ratio of perplexity on low and high ads grows from 2.79 to 3.01.

Qualitative Results: Figs. 3-6 and Listings 1-10 show some randomly sampled ad storyboards generated by Henry-SEED and Sec. D.1 contains some expert comments over the generated ad storyboards. These qualitative examples are generated by prompting Adobe Firefly [1] with the scene descriptions provided by Henry-SEED[7], followed by pasting OCR from the Henry-SEED generated verbalization on top of the generated images. We provide visualizations for easier understanding (Figs. 3-6), along with the raw generations (Listings 1-10). We also run some ablation studies to find the impact of the amount of data and architecture on memorable ad generation. We discuss the results in Appendix:§C.2.

5. Conclusion

In this work, we presented the first large-scale ad memorability study and dataset, LAMBDA, measuring long-term memorability. Despite the importance that advertising plays in day-to-day, no large-scale works have tried to model long-term memorability on this multimodal content type. We then presented our model, Henry, which incorporates world and cognitive knowledge to understand the semantics of the ad content, brand, and experimental protocol, ultimately consolidating them together to predict memorability. Henry, when tested on eight datasets across the literature, spanning both short-term and long-term memorability, gets state-of-the-art performance on all of them. Next, we propose the task of generating memorable ads and release a large scale dataset UltraLAMBDA, consisting of 5 million ads for this task. We propose a new method based on self-rewarding language model to generate more memorable ads, which we call, SEED. Finetuning Henry using SEED results in an improvement of over 44% in content memorability.

Acknowledgements: Rajiv Ratn Shah is partly supported by the Infosys Center for AI, the Center of Design and New Media, and the Center of Excellence in Healthcare at IIIT Delhi.

[7]Note: We do not make any changes to Henry-SEED's generation for the voice-over or the scene descriptions before passing it to Firefly.

References

[1] Adobe. Adobe Firefly. https://www.adobe.com/products/firefly.html, 2024. Accessed: February 9, 2024. 8, 13

[2] Erdem Akagunduz, Adrian G Bors, and Karla K Evans. Defining image memorability using the visual memory schema. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2165–2178, 2019. 1

[3] Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 6

[4] A Selin Atalay, Siham El Kihal, and Florian Ellsaesser. Creating effective marketing messages through moderately surprising syntax. *Journal of Marketing*, page 00222429231153582, 2023. 2

[5] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier, 1968. 2

[6] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 6

[7] Steven Bellman, Shruthi Arismendez, and Duane Varan. Can muted video advertising be as effective as video advertising with sound? *SN Business & Economics*, 1(1):27, 2021. 2

[8] Aanisha Bhattacharyya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. A video is worth 4096 tokens: Verbalize videos to understand them in zero shot. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9822–9839, Singapore, Dec. 2023. Association for Computational Linguistics. 6

[9] Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024. 22

[10] Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40, 1992. 22

[11] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 8

[12] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015. 1

[13] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 6

[14] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016. 25

[15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6

[16] Romain Cohendet, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engilberge. Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2531–2540, 2019. 1, 2, 4, 5

[17] Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. You had me at hello: How phrasing affects memorability. In Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park, editors, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 892–901, Jeju Island, Korea, July 2012. Association for Computational Linguistics. 6

[18] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022. 25

[19] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020. 19

[20] Théo Dumont, Juan Segundo Hevia, and Camilo L. Fosco. Modular memorability: Tiered representations for video memorability prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10751–10760, June 2023. 5, 20

[21] Hermann Ebbinghaus. *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot, 1885. 2

[22] Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Amnet: Memorability estimation with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6363–6372, 2018. 5

[23] Forbes. Agencies agree 2021 was a record year for ad spending, with more growth expected in 2022. https://www.forbes.com/sites/bradadgate/2021/12/08/agencies-agree-2021-was-a-record-year-for-ad-spending-with-more-growth-expected-in-2022/, 2022. Accessed on December 8, 2023. 1

[24] Doron Friedman and Yishai A Feldman. Knowledge-based cinematography and its applications. In *ECAI*, volume 16, page 256. Citeseer, 2004. 6

[25] Theodoros Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12), 2015. 19

[26] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 5744–5753, 2019. 6

[27] Lore Goetschalckx and Johan Wagemans. Memcat: a new category-based image set quantified on memorability. *PeerJ*, 7:e8169, 2019. 2

[28] Thomas Hagen and Thomas Espeseth. Image memorability prediction with vision transformers. *arXiv preprint arXiv:2301.08647*, 2023. 5, 6, 20

[29] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1469–1482, 2013. 7

[30] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR 2011*, pages 145–152. IEEE, 2011. 1, 2, 5

[31] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 08 2005. 22

[32] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. 7

[33] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10041–10051, 2023. 19

[34] Ashmit Khandelwal, Aditya Agrawal, Aanisha Bhattacharyya, Yaman K Singla, Somesh Singh, Uttaran Bhattacharya, Ishita Dasgupta, Stefano Petrangeli, Rajiv Ratn Shah, Changyou Chen, et al. Large content and behavior models to understand, simulate, and optimize content and behavior. *The Journal of Machine Learning Research*, 2024. 3

[35] Aditya Khosla, Wilma A Bainbridge, Antonio Torralba, and Aude Oliva. Modifying the memorability of face photographs. In *Proceedings of the IEEE international conference on computer vision*, pages 3200–3207, 2013. 6

[36] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, pages 2390–2398, 2015. 1, 2, 5

[37] Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. Synthesizing human gaze feedback for improved nlp performance. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1887–1900, 2023. 19

[38] Rukiye Savran Kiziltepe, Lorin Sweeney, Mihai Gabriel Constantin, Faiyaz Doctor, Alba García Seco de Herrera, Claire-Héléne Demarty, Graham Healy, Bogdan Ionescu, and Alan F. Smeaton. An annotated video dataset for computing video memorability. *Data in Brief*, 39:107671, dec 2021. 2

[39] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 6

[40] Yaman Kumar, Rajat Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. Persuasion strategies in advertisements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 57–66, 2023. 6

[41] Cameron Kyle-Davidson, Adrian G Bors, and Karla K Evans. Modulating human memory for complex scenes with artificially generated images. *Scientific Reports*, 12(1):1583, 2022. 6

[42] Robert J Lavidge and Gary A Steiner. A model for predictive measurements of advertising effectiveness. *Journal of marketing*, 25(6):59–62, 1961. 1, 2

[43] Cong Li. Primacy effect or recency effect? a long-term memory test of super bowl commercials. *Journal of Consumer Behaviour: An International Research Review*, 9(1):32–44, 2010. 2

[44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3, 6, 19

[45] Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *International Conference on Learning Representations*, 2022. 6

[46] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 19

[47] Youwei Lu and Xiaoyu Wu. Cross-modal interaction for video memorability prediction. In Steven Hicks, Konstantin Pogorelov, Andreas Lommatzsch, Alba García Seco de Herrera, Pierre-Etienne Martin, Syed Zohaib Hassan, Alastair Porter, Asem Kasem, Stelios Andreadis, Mathias Lux, Marc Gallofré Ocaña, Alex Liu, and Martha A. Larson, editors, *Working Notes Proceedings of the MediaEval 2021 Workshop, Online, 13-15 December 2021*, volume 3181 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. 5

[48] Li-Wei Mai and Georgia Schoeller. Emotions, attitudes and memorability associated with tv commercials. *Journal of Targeting, Measurement and Analysis for Marketing*, 17:55–63, 2009. 2

[49] Marta Malavolta, Emiliano Trimarco, Vida Groznik, and Aleksander Sadikov. Awareness of being tested and its effect on reading behaviour. In *International Conference on Artificial Intelligence in Medicine*, pages 365–370. Springer, 2022. 2

[50] Rob McCarney, James Warner, Steve Iliffe, Robbert Van Haselen, Mark Griffin, and Peter Fisher. The hawthorne effect: a randomised, controlled trial. *BMC medical research methodology*, 7(1):1–8, 2007. 2

[51] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. 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*, pages 223–240. Springer, 2020. 2, 4, 5

[52] Kate Newstead and Jenni Romaniuk. Cost per second: The relative effectiveness of 15-and 30-second television advertisements. *Journal of Advertising Research*, 50(1):68–76, 2010. 2

[53] Dennis Norris. Short-term memory and long-term memory are still different. *Psychological bulletin*, 143(9):992, 2017. 2, 19

[54] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 6

[55] Sanjay Putrevu, Joni Tan, and Kenneth R Lord. Consumer responses to complex advertisements: The moderating role of need for cognition, knowledge, and gender. *Journal of Current Issues & Research in Advertising*, 26(1):9–24, 2004. 2

[56] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. volume 106, page 107404, 2020. 19

[57] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. 19

[58] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. 25

[59] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. 25

[60] Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. {ZeRO-Offload}: Democratizing {Billion-Scale} model training. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 551–564, 2021. 25

[61] Fritz Jules Roethlisberger and William J Dickson. *Management and the Worker*, volume 5. Psychology press, 2003. 2

[62] Susanne Schmidt and Martin Eisend. Advertising repetition: A meta-analysis on effective frequency in advertising. *Journal of Advertising*, 44(4):415–428, 2015. 2

[63] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image

captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 6

[64] Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. How to make an image more memorable? a deep style transfer approach. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 322–329, 2017. 6

[65] Somesh Singh, Harini S I, Yaman K Singla, and Balaji Krishnamurthy. Images and videos to detect emotions through natural language intermediary. *arXiv preprint*, 2023. 19

[66] Somesh Singh, Harini SI, Yaman K Singla, Veeky Baths, Rajiv Ratn Shah, Changyou Chen, and Balaji Krishnamurthy. Llava finds free lunch: Teaching human behavior improves content understanding abilities of llms. *arXiv preprint arXiv:2405.00942*, 2024. 6

[67] Larry R Squire. The legacy of patient hm for neuroscience. *Neuron*, 61(1):6–9, 2009. 2

[68] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3, 6

[69] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3, 6

[70] Duane Varan, Magda Nenycz-Thiel, Rachel Kennedy, and Steven Bellman. The effects of commercial length on advertising impact: What short advertisements can and cannot deliver. *Journal of Advertising Research*, 60(1):54–70, 2020. 2

[71] Nancy C Waugh and Donald A Norman. Primary memory. *Psychological review*, 72(2):89, 1965. 2

[72] Wikipedia contributors. Sensitivity index. https://en.wikipedia.org/wiki/Sensitivity_index, 2024. 5

[73] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching, 2022. 6

[74] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 19

[75] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3