

Long-Term Ad Memorability: Understanding & Generating Memorable Ads

Harini S I^{*} Somesh Singh^{*} Yaman Kumar^{*} 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

Marketers spend billions of dollars on advertisements but to what end? At the time of purchase, if customers cannot recognize the brand for which they saw an ad, the money spent on the ad is essentially wasted. Despite its importance in marketing, until now, there has been no study on the memorability of ads in the ML literature. Most studies have been conducted on short-term recall (<5 mins) on specific content types like object and action videos. On the other hand, the advertising industry only cares about long-term memorability (a few hours or longer), and ads are almost always highly multimodal, depicting a story through its different modalities (text, images, and videos). With this motivation, we release the first large-scale memorability dataset, LAMDBA, 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 - both content and human factors. For *e.g.*, we find that brands that use commercials with fast-moving scenes are more memorable than those with slower scenes ($p=8e-10$) and that people who use ad-blockers remember a lower number of ads than those who don't ($p=5e-3$). Next, with the motivation of simulating the memorability of marketing materials for a particular audience, ultimately helping create one, we present a novel model, Henry, trained to leverage real-world knowledge of LLMs and visual knowledge of visual encoders to predict the memorability of a content. We test Henry on all the prominent memorability datasets in literature (both images and videos) and achieve state-of-the-art performance across *all* of them. Henry shows strong generalization performance showing better results in 0-shot on unseen datasets. Next, we propose the task of memorable ad generation and release a large-scale ad dataset, UltraLAMDBA, consisting of 4 million ads with their Henry-assigned memorability scores. We show that aligning Henry to generate memorable content results in the improvement of memorability scores by more than 25%.

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 (Forbes, 2022). 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 (Lavidge and Steiner, 1961). 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 (Bylinskii et al., 2015). Human factors represent a person’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 (Isola et al., 2011; Khosla et al., 2015; Cohendet et al., 2019) 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 was remembered. On the other hand, high-level features like object and scene semantics have significant correlation with memorability.

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** - Newman et al. (2020), **VideoMem** - Cohendet et al. (2019), **LaMem** - Khosla et al. (2015), **SUN** - Isola et al. (2011), **MemCat** - Goetschalckx and Wagemans (2019), **MediaEval** - Kiziltepe et al. (2021)

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

bility. For example, images with people are more memorable than those without people. Similarly, images with outdoor scenes are less memorable than those with indoor scenes. Further, these initial studies contributed to protocols for conducting memorability studies. They proposed a memorability game where they played a competitive 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 limited the scope of these studies to short-term memorability (few seconds to few minutes), and the competitive nature made 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.

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 (Lavidge and Steiner, 1961). Due to the purchase funnel being a multi-stage lengthy process, long-term memorability (LTM) is the closest proxy to model customer memory (Norris, 2017; Waugh and Norman, 1965). While the LTM store (as distinct from the STM store) has been studied for over five decades in psychology (Ebbinghaus, 1885; Atkinson and Shiffrin, 1968), there have been no large-scale datasets that can help us train a model for customer LTM. 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 game further disconnects the STM 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 (2010) conducted a field experiment on advertisements shown during the 2006 Super Bowl Games where they asked Super Bowl viewers 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 (Atalay et al., 2023), emotional content (Putrevu et al., 2004; Mai and Schoeller, 2009), repetition (Schmidt and Eisend, 2015), advertisement spot length (Newstead and Romaniuk, 2010; Varan et al., 2020), brand logo and imagery at a prominent position in the ad (early branding) (Newstead and Romaniuk, 2010), presence of sound (Bellman et al., 2021), and also on customer factors like involvement and relevance (Newstead and Romaniuk, 2010; Schmidt and Eisend, 2015).

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 limited number of advertisements and are unsuitable for training ML models for LTM simulation. While there have been many works in the ML literature itself to model STM (Isola et al., 2011; Cohendet et al., 2019), due to this lack of large-scale datasets for LTM, there have been no works in the ML literature to model LTM.

Therefore, to model long-term memory over advertisements, we conduct the first large-scale human study on long-term advertisement memorability¹. We call it LAMBDA (Long-term Ad MemoraBility

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

DATASET). To collect LAMBDA, we conduct long-term memorability study involving 1749 participants across four sessions across two institutes. 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? In order to model customer memory, we design a novel architecture, Henry² (Fig. 2), incorporating world-knowledge from large language models (Llama (Touvron et al., 2023)) and visual knowledge from vision encoders (EVA-CLIP (Sun et al., 2023)). 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 (Li et al., 2023a) and further augment them with scene “verbalizations” (Bhattacharyya et al., 2023) involving scene descriptors like visual caption, optical character recognition (OCR), automatic speech recognition (ASR), emotion, color distribution, and scene complexity scores, which help the model ground the visual knowledge in the LLM’s world knowledge. We train the model on our long-term memorability 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 (Khosla et al., 2015), MemCat (Goetschalckx and Wagemans, 2019), SUN (Isola et al., 2011), Memento10k (Newman et al., 2020), MediaEval (Kiziltepe et al., 2021), and obtain state-of-the-art performance on all of them. Further, we show that Henry performs well on unseen datasets in zero-shot settings getting better performance than models trained on those datasets.

How do we generate memorable Ads? One of the primary goals of modeling memorability on ads is to generate more memorable ads. We formulate the task of generating more memorable ads as given the advertisement description (also called marketing brief)

containing brand and campaign name, to generate the ad scenes and dialogues. Drawing inspiration from advancements in reinforcement learning from human feedback (RLHF) research (Ouyang et al., 2022; Stiennon et al., 2020), which addresses the generation of less harmful content, we adopt a two-step process: instruction fine-tuning and preference alignment. To accomplish the generation of memorable ads, we collect the first large-scale advertisements dataset, collecting brand name, ad text, time, ad content, and channel from the web. We use Henry trained in the previous step to simulate memorability on the collected samples, ultimately getting a dataset of 4 million advertisements with their automatic speech transcripts, OCR, automatically detected objects, colors, aesthetics, captions, emotions, logos, and memorability scores. We call this dataset UltraLAMBDA. Subsequently, we employ UltraLAMBDA to further train Henry using implicit language q-learning (Snell et al., 2022), aiming to generate ads with heightened memorability resulting in an improvement of over 25% memorability score.

We make the following contributions:

1. We release the first large-scale dataset, LAMBDA, on long-term advertisement memorability involving more than 1700 participants across four sessions conducted in two institutes. 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).
2. We design a novel model, Henry, which can model both short-term and long-term memorability and can incorporate scene understanding, brand knowledge, and speech (§3). Henry achieves state-of-the-art performance on eight image and video memorability datasets in the literature (§3.5). Further, we show that Henry performs well on unseen datasets in zero-shot settings getting better performance than models trained on those datasets.
3. Next, we propose the task of memorable ad generation. We release the first large scale ad dataset, UltraLAMBDA, consisting of 4 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, using ILQL, we train Henry to progressively generate more memorable ads resulting an improvement of 25% in memorability scores (§4).
4. We conduct an extensive set of experiments on memorability prediction, showing the effects of LTM

²We name the model Henry in honor of the immense contributions by the patient Henry Molaison (H.M.) (Squire, 2009). An experimental surgery conducted on him resulted in the discovery of the distinct regions responsible for LTM and STM store.

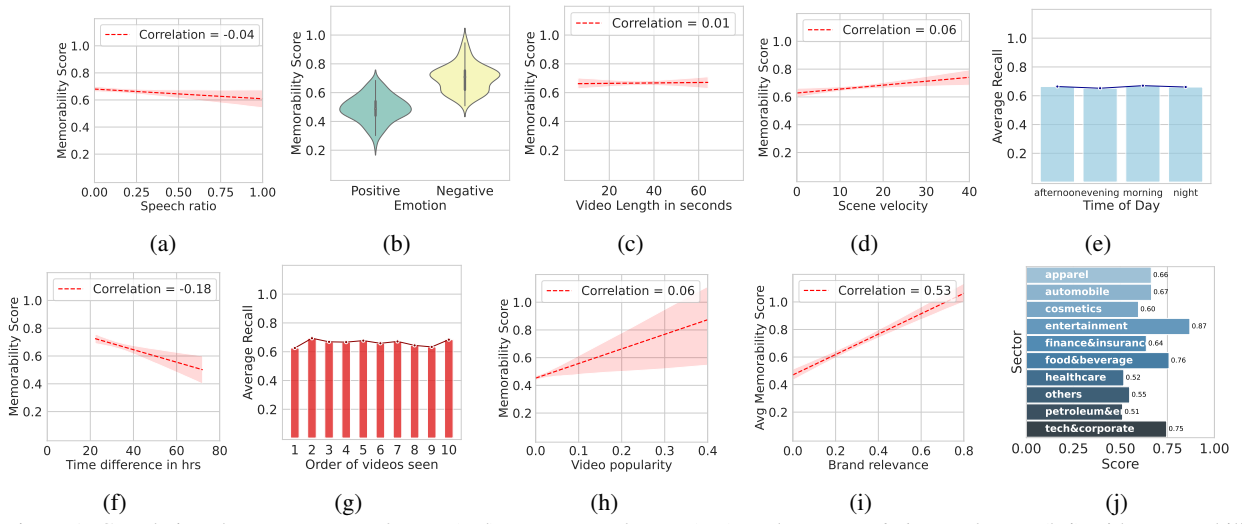


Figure 1: Correlations between *content factors* (a-d), *interaction factors* (e-g), and *customer behavior factors* (h-j) with memorability on LAMBDA samples. While emotion has a high correlation with memory, other content factors do not have much correlation. Further, while there is little correlation 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.

on STM modeling and vice-versa, and the effects of changing world-knowledge with time, scene understanding, brand knowledge, and speech on memorability modeling (§3.5).

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, mostly of actions by humans, and do not depend on any prior context (Newman et al., 2020; Cohendet et al., 2019), the videos in our dataset come from multimodal advertisements released on Youtube channels of 276 major brands covering 113 industries³. We collect 2205 such advertisements 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, visual and 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 initial questionnaire, participants proceed to the core segment of the study, where they are shown 11 advertisements in a sequential manner. Notably, the eleventh advertisement is deliberately repeated for half of the participants, while it is unique for the other half. After the 11th video, participants are asked if they recollect watching it in the span of the study. 57% participants are able to recognize the repeated video correctly. To ensure participant engagement and attentiveness throughout the study, attention-check questions are placed between every two or three advertisements. These questions are simple common sense questions like “How many legs does a cow have?”. If the participant fails to answer the question within 10 secs, they are made to rewatch the previous ad.

Next, we test their memorability over the span of the next 1-3 days. We assess two things: brand recognition and ad recall. For the former, we present participants with a list of 20 options tasking them with identifying brands they remember encountered during the previous session. For the latter, participants

³Industry information was collected from Wikidata entry of the brand.


Image	Semantic Category	Verbalization	Semantic Category	Verbalization
	OCR	The text shown in the scene is "Adidas".	Clutter	The clutter in the scene is low .
	ASR	The audio in the scene is "To take hold of the world's spotlight overnight".	Photo Style	The photography style of the scene is commercial photography .
	Human Presence	The scene has 1 person with prominent face.	Emotion	The emotion of the scene is ambitious, determined .
	Caption	The scene shows a young woman sitting in a glass door looking out.	Aesthetics	The image has medium aesthetic value.
	Colors	The foreground colors of the scene are Black, Dark Brown, Dark Blue, Dark Gray, Mud Green and the background colors are Dark Blue, Black, Dark Brown . The dominant tone of the scene is neutral .	Object Tags	This scene is categorized by the tags: person, woman, blazer, facing, template, fashion, street fashion, cold, client, cardigan, sweat .
	Audio Type	The scene has music and speech .	Logo Presence	There is a logo in the scene.

Table 2: To augment the scene understanding of LLM, we verbalize video scenes and images using a diverse set of perception tools and pass it to the LLM in the format shown in the table. For image memorability datasets, we use the following semantic categories: caption, color, photo style, emotion, clutter, human presence, object tags, OCR, and aesthetics. For video scene memorability datasets, we use the following semantic categories: caption, color, emotion, human presence, object tags, ASR, OCR, Audio-type, Logo-presence. We use the following models to extract the features: OCR (Du et al., 2020), clutter (Khurana et al., 2023), ASR (Radford et al., 2022), Photo style (Li et al., 2023a), human presence (Liu et al., 2023b), emotion (Singh et al., 2023), caption (Li et al., 2023a), aesthetics (Ke et al., 2023), colors (Qin et al., 2020), object tags (Zhang et al., 2023b), audio-type (Giannakopoulos, 2015), and logo presence (Zhang et al., 2023b). Black colored text is the verbalization template, and red text indicates the model outputs.

are asked to describe what they remember about the ads of the recognized brands⁴. 971 participants took the memorability test in a take-home setting, and the other 778 took the test together in an auditorium.

2.3 What makes an Ad memorable?

Among the many reasons why an ad might be memorable, we investigate first 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 as measured by average participant ratings, video popularity as measured by Youtube likes).

Content Factors: To answer the question of what is in the content which determines memory, previous studies (Isola et al., 2011; Newman et al., 2020) 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. (2020) 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 has very little correlation with 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, in line with other studies (Newstead and Romaniuk, 2010; Varan et al., 2020), 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. We show the effect of time of day on memorability in Fig. 1e. We see that time of day of watching has almost no effect on the memorability of the ad. Further, it may be expected that memorability decays as time passes. Comparing forgetting curve for ads (Fig. 1f) with action videos (Cohendet et al., 2019), we see that ad videos have a similar forgetting curve as action videos but a different forgetting coefficient. The difference in forgetting coefficient most likely arises due to differences in protocols. They used a two-stage memory protocol in which participants did both short-term and 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. This is in line with other results on ad memorability (Terry, 2005).

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

⁴The complete questionnaire for participant one is given in Appendix:§8.

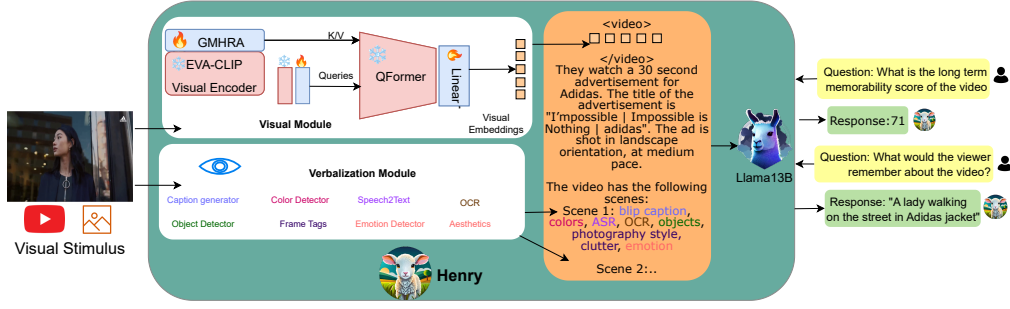


Figure 2: Predicting memorability by encoding visual concepts captured through visual encoder (EVA-CLIP) and world knowledge captured through LLM (Llama). To leverage the rich knowledge of LLMs, we use GMHRA and QFormer to convert visual tokens of ViT to language tokens which Llama can understand. Further, we find that verbalizing the visual stimulus helps Llama to gather information more explicitly than what is provided by ViT+QFormer. 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.

personally used from a set of 15 randomly chosen brands and similarly choose brands they have seen ads for. In order 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 which needs (a) *visual knowledge* to identify and understand visual concepts across images and videos like shapes, colors, objects, and scenes, and also (b) *world knowledge* to relate the captured visual concepts to real-world concepts capturing their function, use, and interaction patterns. Recently, models like BLIP (Li et al., 2023a), Llava (Liu et al., 2023a), and Video4096 (Bhattacharyya et al., 2023) have shown that visual signals can be successfully mapped to language semantics, allowing us to leverage visual knowledge from visual embedding models like ViT (Dosovitskiy et al., 2020) and world knowledge from LLMs. For instance, when Airbnb⁵ shows an adult female and a male with the text, “Our guest room is paying for our wedding”, most likely denotes a couple saying that renting out their space on Airbnb helps them sponsor their wedding (Kumar et al., 2023). World knowledge captured in LLMs, together with the visual knowledge of ViT, helps to identify the two adults as a couple and relate the text with the Airbnb logo and make sense of all three concepts together.

Fig. 2 shows our architecture to predict user memorability. The overall process involves passing visual

content in the form of video or image through a visual embedder, which maps the visual content into a language space with the help of QFormer (Li et al., 2023a). We find that describing the visual content in language and combining it with the video/image tokens helps for better prediction. The resulting content tokens, along with experiment descriptions, are fed into the Henry to predict memorability scores between 00 to 99. We discuss each part of the architecture and training next.

3.1 Encoding Visual Content

The primary goal of this step is to effectively leverage the “world-knowledge” capabilities of the pre-trained LLM. We choose Llama (Touvron et al., 2023) as our base LLM. We employ two techniques to convert visual data into language: encoding visual frames into the LLM space and verbalizing visual content into language space. We detail these two steps next:

Sampling Frames: While images only have one frame and can directly be passed to the encoding stage, videos need some preprocessing before they can be encoded. To handle videos, we sample their most important frames. We try a sampling technique inspired by cinematography principles and involves converting frames to the HSV color space. 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 (Arev et al., 2014; Friedman and Feldman, 2004). 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 Xu et al. (2022).

Encoding Into Language Embedding Space: Henry employs the EVA-CLIP visual embedder (Sun et al., 2023) to encode visual content. While ViT captures the spatial interactions in an image frame well, Global Multi-Head Relation Aggregator (GMHRA)

⁵see Appendix Fig. 5) for the ad

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 (ViT-Mem)	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 on individual datasets	0.74	0.82	0.73	-	0.75	0.64	0.50	0.55
Henry on combined datasets	0.72	0.79	0.76	0.79	0.72	0.60	0.48	0.52

Table 3: 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. Literature SOTA Models (dataset: SOTA): LaMem: [Hagen and Espeseth \(2023\)](#), MemCat: [Hagen and Espeseth \(2023\)](#), SUN: [Fajtl et al. \(2018\)](#), Merged Image datasets: [Hagen and Espeseth \(2023\)](#), Memento10k: [Dumont et al. \(2023\)](#), VideoMem: [Dumont et al. \(2023\)](#), MediaEval: [Lu and Wu \(2021\)](#)

from UniFormer ([Li et al., 2022](#)) helps aggregate the information better across the time dimension. The combination of ViT and GMHRA gives us a good representation for visual content. Next, to effectively leverage the LLM’s rich language representations, we use Q-Former from BLIP-2 ([Li et al., 2023a](#)) with an extra linear layer and additional query tokens to convert from visual tokens to language tokens.

3.2 Verbalizing Visual Content

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 semantic information present in images (Table 6). Therefore, to augment the visual understanding of the LLM, we verbalize the frame semantic information using a diverse set of perception tools (Table 2). These perception tools help ground the visual perception of LLM in the semantic content of the image. Specifically, we use the following information to understand scenes and images better (Table 2): image caption, photography style, visual emotions, clutter, human presence, OCR (Optical Character Recognition), image aesthetics, visual tags, orientation, foreground/background colors, and tone. For videos, we use the following additional information: YouTube provided title and description, video duration, and brand information. We pass the video story, described by the concatenation of scene descriptions similar to images, along with their corresponding timestamps and flow based scene velocity to the LLM (Appendix: Listing 3).

3.3 Verbalizing Experiment Conditions

We find that verbalizing experimental conditions help give context to the LLM about the task. Details like task type (long-term, short-term), data distribution (for eg: video advertisements, images of natural scenes), subject descriptions (for eg: college students from X college) are given to the LLM. For example, the subjects are students from MIT, and they are shown images of faces and their short term memorability is recorded. We find that this also helps us in solving another challenge: a few samples in datasets like Lamem and Memcat are repeated, but since the

experiments were conducted in different settings, the model gets confused if we don’t explain the experimental context.

3.4 Two-Stage Training

Following prior works like ([Liu et al., 2023a](#); [Li et al., 2023b,a](#); [Zhu et al., 2023](#); [Ge et al., 2023](#); [Zhang et al., 2023a](#)), we do a two-stage training paradigm where in the first stage, we utilize the Webvid ([Bain et al., 2021](#)), COCO caption ([Chen et al., 2015](#)), Visual Genome ([Krishna et al., 2017](#)), CC3M ([Sharma et al., 2018](#)), and CC12M ([Changpinyo et al., 2021](#)) datasets to align the visual encoder embeddings with LLM via a large-scale pretraining 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. 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.

3.5 Results and Discussion

We conduct extensive experiments on eight datasets, covering both videos and images, STM and LTM. Table 3 shows the results of Henry⁶ compared with the current state-of-the-art models in the literature. We find that Henry outperforms all models in the literature across all the datasets. Further, we compare Henry with 10-shot GPT-3.5 (text-davinci-003) ([Ouyang et al., 2022](#)) as well, where we provide GPT with the same verbalization (for 10 examples), as we provided to Henry. We also compare the performance with prior regression based methods using features extracted from ViT L-14 ([Hagen and Espeseth, 2023](#)).

Next, to understand effect of different kinds of data and architectural choices, we conduct two types of

⁶Computing infrastructure used to conduct the experiments along with hyperparameters are given in Appendix:§10.1. All experiments are conducted with three random seeds and averages are reported.

Model	Training	Memorability Reward		Preference
		LM	HM	
GPT 3.5	5-shot	0.38	0.82	-
GPT 4	5-shot	0.39	0.81	-
GPT 3.5	3-shot	0.34	0.75	-
Henry	3-shot	0.29	0.62	12%
Henry	SFT	0.34	0.6	41%
Henry	PFT	0.40	0.77	34%
Henry	SFT+RLHF	0.48	0.83	39%
Henry	PFT+RLHF	0.54	0.88	35%
Test Data	-	0.41	0.89	84%

Table 4: Results of Henry compared with in-context-learning GPT-3.5, 4 on Memorability Reward and Preference. Memorability reward is computed on ground-truth LAMBDA samples (LM = samples with memorability < 0.75 , HM = samples with memorability > 0.75). Preference denotes how much % times GPT-3.5 prefers the generated scenes and dialogues. Henry is able to improve memorability on LM samples by 25%. Interestingly, the content generated by GPT-3.5 and 4 has low memorability.

ablations. Table 5 shows the data ablations. We see that STM helps in predicting LTM relatively much better than vice versa. Studies in psychology show that for a content to get committed to LTM, it has to pass through STM (Norris, 2017). Therefore, content memorable, according to STM, has an effect on LTM but, interestingly, not vice versa. Further, we observe that Henry loses performance for unseen brands. This underscores the importance of scaling the study across more brands. Next, we evaluate the impact of various architectural choices (Table 6). We find that Henry’s vision branch is not strong enough by itself to produce good results. Interestingly, lower-level features like objects and colors have the maximum impact on STM, but higher-level features like emotion, ASR, and aesthetics have a higher impact on LTM.

4 Generating Memorable Ads

We propose the new task of memorable ad generation. Given the input as brand name and a brief campaign description, the task is to generate an ad consisting of memorable scenes and dialogues. Most of the work in memorability has been about *how much* a content is memorable; there is a small amount of work to generate memorable content (Khosla et al., 2013; Siarohin et al., 2017; Goetschalckx et al., 2019; Danescu-Niculescu-Mizil et al., 2012), and most of it is in the computer vision domain. Therefore, with the aim of training models to generate memorable ads, we collect a dataset consisting of 4 million ads sourcing the dataset from the web. We collect the brand name, ad title, posted caption, date, and raw ad content (video and images). We automatically label the collected ad content for ASR, OCR, colors, aesthetics, emotions, and memorability scores. We call this dataset UltraLAMBDA.

Methodology: First, to teach Henry to generate ads

given the brand name and marketing brief, we fine-tune it on the UltraLAMBDA samples (Fig. 3). We define two dataset splits for this task Supervised-Fine-Tuning (SFT): fine-tuning on all (potentially noisy) samples of UltraLambada, Preference-Fine-Tuning (PFT): SFT combined with oversampling of (cleaner) most memorable ads from LAMBDA. Since we only generate the verbalization, not the actual video rendition, we consider offline learning RLHF algorithm, ILQL (Snell et al., 2022). We compare generations of PFT, SFT, PFT+RLHF, and SFT+RLHF models with in-context-learning based Henry, GPT-3.5, 4 generated ads.

To evaluate, we use two metrics, *GPT-3.5 preference* to measure the overall generation quality and *Memorability Reward* to estimate the memorability of the generated ads. *GPT-3.5 Preference* is the percentage of times GPT-3.5 prefers the generated ad compared to its own generation (similar to vicuna eval (Chiang et al., 2023)). This measures the overall quality of the generated advertisement. *Memorability Reward* is the mean predicted memorability using Henry while keeping its visual branch masked (because the rendition of the generated video is not available). Table-6 shows that masking the visual modality still gives reasonable estimates for memorability.

Results: To analyze the results, we divide the test data into two parts Low Memorability (LM) i.e. LAMBDA videos with < 0.75 memorability score, and remaining High Memorability (HM). We observe that Henry PFT+RLHF, while retaining the memorability of samples in the HM bucket, can increase the Memorability Reward by 25% for LM. We also notice that the generation quality is not correlated with the memorability reward. See §7 for some example generations. We observe that Henry puts more emphasis on including brand names and emotions in the ad.

5 Conclusion

In this work, we presented the first large-scale ad memorability 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 knowledge and visual 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 Ultra-LAMBDA for this purpose. Finetuning and aligning Henry for this task results in an improvement of over 25% in content memorability.

6 Limitations

In this paper, we try to fill a gap in the existing literature about long-term memorability modeling and datasets. Therefore, we conduct the first study for that purpose. While doing that, we have made initial efforts starting with the English language advertisements. Future work would be needed to address other languages. Further, given the limitations of the study, we conducted it in an academic environment with the student population consisting of undergraduate and graduate student volunteers. We will expand the scope to a wider audience in the future work. We trained a model Henry on the collected dataset, showing good performance on all literature datasets. However, since the literature datasets are all English based and deal with majorly uniform population, scaling the training to more languages and population types will be taken up in the future work. We also observed a decrease in performance for brands not seen during the training and for videos with longer verbalizations exceeding 1500 tokens. Additionally, the model exhibits a slight inaccuracy when advertisements have significant musical content. Other than these limitations, in our best opinion, the model does not pose any potential risk or harm.

References

- Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. 2014. Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics (TOG)*, 33(4):1–11.
- A Selin Atalay, Siham El Kihal, and Florian Ellsaesser. 2023. Creating effective marketing messages through moderately surprising syntax. *Journal of Marketing*, page 00222429231153582.
- Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier.
- Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. 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.
- Steven Bellman, Shruthi Arismendez, and Duane Varan. 2021. Can muted video advertising be as effective as video advertising with sound? *SN Business & Economics*, 1(1):27.
- Aanisha Bhattacharyya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. 2023. [A video is worth 4096 tokens: Verbalize videos to understand them in zero shot](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9822–9839, Singapore. Association for Computational Linguistics.
- Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. 2015. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178.
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. 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.
- Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality](#).
- Romain Cohendet, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engilberge. 2019. 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.
- Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. [You had me at hello: How phrasing affects memorability](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 892–901, Jeju Island, Korea. Association for Computational Linguistics.
- Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. 2020. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*.
- Théo Dumont, Juan Segundo Hevia, and Camilo L. Fosco. 2023. 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.
- Hermann Ebbinghaus. 1885. *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot.
- Jiri Fajtl, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. Amnet: Memorability estimation with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6363–6372.

- Forbes. 2022. 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/>. Accessed on December 8, 2023.
- Doron Friedman and Yishai A Feldman. 2004. Knowledge-based cinematography and its applications. In *ECAI*, volume 16, page 256. Citeseer.
- Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*.
- Theodoros Giannakopoulos. 2015. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12).
- Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. 2019. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5744–5753.
- Lore Goetschalckx and Johan Wagemans. 2019. Memcat: a new category-based image set quantified on memorability. *PeerJ*, 7:e8169.
- Thomas Hagen and Thomas Espeseth. 2023. Image memorability prediction with vision transformers. *arXiv preprint arXiv:2301.08647*.
- Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable? In *CVPR 2011*, pages 145–152. IEEE.
- Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. 2023. 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.
- Aditya Khosla, Wilma A Bainbridge, Antonio Torralba, and Aude Oliva. 2013. Modifying the memorability of face photographs. In *Proceedings of the IEEE international conference on computer vision*, pages 3200–3207.
- Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, pages 2390–2398.
- Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. 2023. 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.
- Rukiye Savran Kiziltepe, Lorin Sweeney, Mihai Gabriel Constantin, Faiyaz Doctor, Alba García Seco de Herrera, Claire-Hélène Demarty, Graham Healy, Bogdan Ionescu, and Alan F. Smeaton. 2021. An annotated video dataset for computing video memorability. *Data in Brief*, 39:107671.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannic Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.
- Yaman Kumar, Rajat Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. 2023. Persuasion strategies in advertisements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 57–66.
- Robert J Lavidge and Gary A Steiner. 1961. A model for predictive measurements of advertising effectiveness. *Journal of marketing*, 25(6):59–62.
- Cong Li. 2010. 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.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.
- Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. 2022. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *International Conference on Learning Representations*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.
- Youwei Lu and Xiaoyu Wu. 2021. Cross-modal interaction for video memorability prediction. In *Working Notes Proceedings of the MediaEval 2021 Workshop, Online, 13-15 December 2021*, volume 3181 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Li-Wei Mai and Georgia Schoeller. 2009. Emotions, attitudes and memorability associated with tv commercials. *Journal of Targeting, Measurement and Analysis for Marketing*, 17:55–63.
- 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*, pages 223–240. Springer.
- Kate Newstead and Jenni Romaniuk. 2010. Cost per second: The relative effectiveness of 15-and 30-second television advertisements. *Journal of Advertising Research*, 50(1):68–76.
- Dennis Norris. 2017. Short-term memory and long-term memory are still different. *Psychological bulletin*, 143(9):992.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Sanjay Putrevu, Joni Tan, and Kenneth R Lord. 2004. 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.

- Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar Zaiane, and Martin Jagersand. 2020. U2-net: Going deeper with nested u-structure for salient object detection. volume 106, page 107404.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#).
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. 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.
- Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. 2021. {ZeRO-Offload}: Democratizing {Billion-Scale} model training. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 551–564.
- Susanne Schmidt and Martin Eisend. 2015. Advertising repetition: A meta-analysis on effective frequency in advertising. *Journal of Advertising*, 44(4):415–428.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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.
- Aliaksandr Siarohin, Gloria Zen, Cveta Majtanovic, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. 2017. 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.
- Somesh Singh, Harini S I, Yaman K Singla, and Balaji Krishnamurthy. 2023. Images and videos to detect emotions through natural language intermediary. *arXiv preprint*.
- Charlie Snell, Ilya Kostrikov, Yi Su, Mengjiao Yang, and Sergey Levine. 2022. Offline rl for natural language generation with implicit language q learning. *arXiv preprint arXiv:2206.11871*.
- Larry R Squire. 2009. The legacy of patient hm for neuroscience. *Neuron*, 61(1):6–9.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.
- Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.
- W Scott Terry. 2005. Serial position effects in recall of television commercials. *The Journal of general psychology*, 132(2):151–164.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Duane Varan, Magda Nenycz-Thiel, Rachel Kennedy, and Steven Bellman. 2020. The effects of commercial length on advertising impact: What short advertisements can and cannot deliver. *Journal of Advertising Research*, 60(1):54–70.
- Nancy C Waugh and Donald A Norman. 1965. Primary memory. *Psychological review*, 72(2):89.
- Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezaatofghi, and Dacheng Tao. 2022. [Gmflow: Learning optical flow via global matching](#).
- Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.
- Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2023b. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix

Generalization Type	Train on	Zero-shot Testing	Lamem	Memcat	SUN	VideoMem	Memento10k	LAMBDA
Memory-type	Short-term	Long-term	-	-	-	0.31	-	0.18
Memory-type	Long-term	Short-term	0.06	0.08	0.07	0.15	0.1	-
Modality	Videos	Images	0.55	0.65	0.55	-	-	-
Modality	Images	Videos	-	-	-	0.44	0.54	0.09
Brands	All except 20 brands	Left-out 20 brands	-	-	-	-	-	0.42
Dataset	All except Memento	Memento	-	-	-	-	0.59	-
Dataset	All except Memcat	Memcat	-	0.68	-	-	-	-

Table 5: Ablation across data to understand how memorability prediction generalizes across the type of memory, datasets, modality (image/video), and brands. The reported values are correlations between model and human memorability scores.

	Lamem	Memcat	VideoMem(ST)	Memento10k	VideoMem(LT)	LAMBDA
Henry on individual datasets	0.74	0.82	0.64	0.75	0.48	0.55
Henry -emotion	0.71	0.78	0.61	0.73	0.42	0.46
Henry -asr	-	-	-	-	-	0.46
Henry vision only	0.20	0.17	0.17	0.21	0.15	0.11
Henry language only	0.51	0.53	0.42	0.54	0.37	0.44
Henry -asr-emotion	-	-	-	-	-	0.42
Henry -object tags	0.67	0.71	0.57	0.69	0.46	0.52
Henry -colors	0.65	0.74	0.55	0.67	0.45	0.51
Henry -aesthetics	0.72	0.79	0.61	0.71	0.46	0.53
Henry -clutter	0.73	0.81	0.60	0.74	0.45	0.53

Table 6: Ablation across architectural choices. “-” denotes non-speech dataset

7 Generated Ad Samples

Below are some examples of voice-over(transcripts) generated by Henry after PFT+ RLHF training.

<p>Listing 1: Voiceover generated by Henry for an Ulta Beauty Ad</p> <p>Original Link to the video ad: https://memorability.blob.core.windows.net/videos/2506_lgJRPZy4ytY.mp4</p> <p>Input Title: How to Avoid Lice with Rosemary Repel by Fairy Tales Ulta Beauty</p> <p>Generated Ad: "Unveil the secret to lice-free bliss with Rosemary Repel by Fairy Tales. Immerse yourself in a tub of serenity, surrounded by products that promise a worry-free experience. Feel the joy as smiles are exchanged, sharing the magic of Fairy Tales in your home. Witness happiness in every jar, as the woman smiles, holding the key to a carefree life. Embrace the enchantment of Ulta Beauty, where beauty meets wellness, and worry fades away."</p>
<p>Listing 2: Voiceover generated by Henry for a Home Depot Ad</p> <p>Original link to the video ad: https://memorability.blob.core.windows.net/videos/1926_6V7nBPul4wE.mp4</p> <p>Input Title: Husky Folding Workbench Garage Storage Ideas</p> <p>Generated Ad: "Transform your garage into more than just a parking space with the Husky Folding Workbench. Crafted with durable steel and a sleek wood top, this workbench is not only sturdy but stylish too. Whether you prefer working sitting or standing, its adjustable height caters to your needs. And when the job is done, it folds flat, saving precious floor space. With a lifetime warranty, it's time to take charge of your workspace. Get smart and tough storage solutions from Home Depot, quality that lasts since 1924."</p>

Listing 3: Verbalization pattern for inputting content and behavior in the same space

<p>Students are shown ads and their memorability is tested after 1 to 3 days. For the given ad:</p> <p><video> .. </video></p> <p>They watch a 15 second advertisement for Chanel.</p> <p>The title of the advertisement is " Comes in Red for a Limited Edition CHANEL Fragrance ".</p> <p>The ad is shot in landscape orientation, at a medium pace.</p> <p>The audio in the ad says: Number 5. Limited Edition. Chanel.</p> <p>Following are the descriptions of each scene:</p> <p>Scene 1:</p> <p>The scene shows a red bottle of perfume that is on a dark surface.</p>

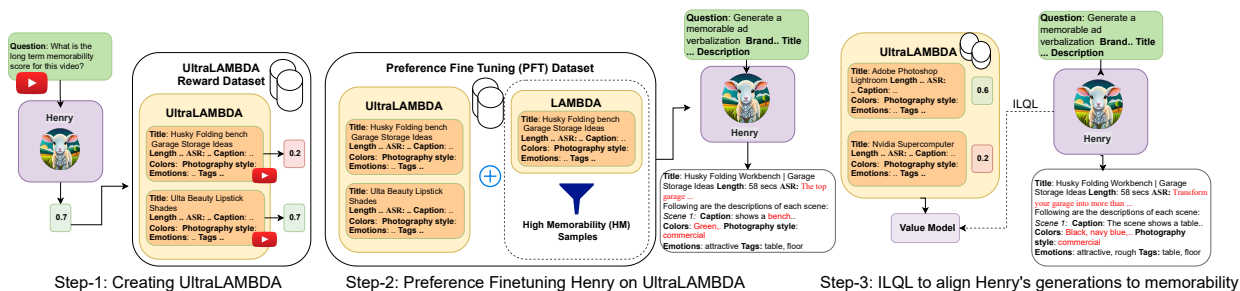


Figure 3: Architecture of Henry for Memorable Ads Generation, this diagram shows three parts 1.UltraLAMBDA Reward Dataset Curation, We use Henry for Memorability simulation to curate a large scale reward dataset. 2. Preference Finetuning, These samples are combined with Filtered high memorability samples for Preference Fine Tuning 3. ILQL for Memorable ad generation using UltraLAMBDA

The foreground colors of the scene are Black, and the background colors are Dark_Brown,Maroon,Black,Gray.
The dominant tone of the scene is neutral.
The photography style of the scene is product.
The scene has Low visual complexity.
The emotions shown in the scene are gift, romantic, celebration.
This scene is categorized by the tags bottle, man, perfume, red, woman.
The text shown in the scene is 'N5', 'CHANEL', 'PARIS', 'PARFUM'

....

What would be the memorability score of this video?

Output: 71

8 Memorability Questionnaire

8.1 Introductory Questionnaire (to be filled before study starts)

- I remember seeing ads for the following brands this year:
 - List 15 randomly selected from the list of brands that we have
- I remember using products of the following brands this year:
 - List 15 randomly selected from the list of brands that we have (non-intersecting list from above)
- Have you installed any Ad Blocking software in your browser(s)?
 - Yes
 - No
- Do you use a Youtube subscription?
 - Yes
 - No
- Approximately how much percentage of time do you spend on Youtube mobile vs Youtube web?
 - <10% on mobile
 - >10% but <30% on mobile
 - >30% but <70% on mobile
 - >70% on mobile
- How do you apprise yourself of the latest products and brands? (Multi correct)
 - Primarily friends and family
 - Amazon, Flipkart or any other e-commerce stores
 - Television and OTT Platform Ads (like Youtube, Netflix, Hotstar, etc)
 - Email Ads

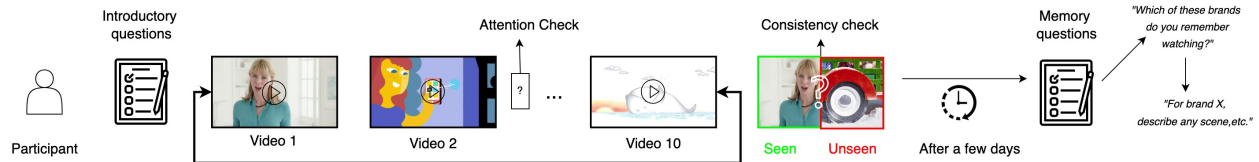


Figure 4: Study protocol

- Store Visits
- Website Ads
- I primarily search for products

8.2 Checks (to be answered during the experiment)

1. **Attention check:** A factual question like, What is the capital of India? (Asked randomly between videos, needs to be answered in <10s)
 - a. Kanpur
 - b. Delhi
 - c. Goa
 - d. Mumbai
2. **Consistency Check:** Do you remember watching this video in this experiment (Asked after showing the 11th video)
 - a. Yes
 - b. No

8.3 Recognition Questions

1. In the study, I remember seeing Ads of the following brands:
 - (Randomly selected list of 20 brands which contains the brands shown to the participant)
 - {For each brand in the list which the participant has selected}
2. Brand: X (already filled in)
 - For the {brand} ad, I remember seeing the following (Write Scene Descriptions, feel free to write any scenes, music, characters, emotions, objects you remember seeing):

9 Annotation Protocol

Figure 4 shows a visualization of the annotation protocol.

9.1 Participant details

The participation was voluntary and the participants were students who were offered optional course credit and freebies like eatables and a chance to see research and know their memorability scores. They were shown a protocol of the study and were required to sign the IRB approval which was prominently displayed. The approval contained details about what kind of data was being collected and how the data would be used. The data collection protocol was approved by the IRB of the participating institution. The aggregate statistics was reported to each candidate after completing the study. Three emails were sent to take-home participants, if they didn't reply within the given time frame, their data was discarded from the experiment.

The participants were primarily graduate and undergraduate students. The participants are from two universities spread across two locations of India. The participants are bilingual and speak a variety of languages, including English. The age range is from 16-35 years and all genders/sexes were encouraged. We saw a roughly 30-70 distribution of females to males.

10 Computing Infrastructure and Hyperparameters

10.1 Simulation

All the experiments were conducted on 8x40 A100 instances. All experiments were performed leveraging DeepSpeed ZeRO stage-3 with cpu offload (Ren et al., 2021; Rasley et al., 2020; Rajbhandari et al., 2020) and Flash-attention (Dao et al., 2022) with gradient-checkpointing (Chen et al., 2016) at bf16 precision. We use AdamW as the optimizer (with fused gelu), the learning rate was kept $2e-5$ for all experiments. The maximum context length for image-only datasets is 500, including public video datasets is 800 and including our dataset is 2048. The corresponding batch sizes are 32,16,8. The gradient accumulation is set to 1 and weight decay is disabled. The warmup steps are set to 20 and residual dropout was kept at 0.25. We train all models for two epochs, but use the checkpoint with best validation spearman correlation.

For all experiments, where we combine datasets, we use a custom sampler to account for dataset imbalance, that ensures a maximum proportion of the dataset in an epoch, here are the maximum proportions. For validation we take 5% of each dataset. We use the provided test splits for public datasets and we use a 15% test split for our dataset

10.1.1 Images

1. **Lamem** 50%
2. **Memcat** 100%
3. **SUN** 100%

10.1.2 Videos

1. **VideoMem** 75%
2. **Memento** 75%
3. **AdsData** 100%
4. **MediaEval** 100%

10.2 Generation

All the experiments were conducted on 8x80 A100 instances. All experiments were performed leveraging DeepSpeed ZeRO stage-2, Flash Attention and Gradient-Checkpointing. $\alpha = 0.001$, $\text{awac_scale} = 1$, $\gamma = 0.99$, $\beta = 0$ $\text{cql_scale} = 0.1$

10.2.1 Inference hyperparameters

$\beta = 4$, $\text{temperature} = 0.8$, $\text{steps_for_target_sync} = 10$, $\tau = 0.7$, $\text{two_qs} = \text{True}$, $\text{lr} = 1e-5$

11 License and Terms of Release

LAMBDA and UltraLAMBDA are sourced from brand videos from Youtube, Facebook Political Ads, and CommonCrawl. The dataset annotations on LAMBDA and UltraLAMBDA will be released under CC BY-NC 4.0 license. The videos themselves are released as per their creators' licenses. The videos or the released data does not contain or disclose any identities of their annotators or any specific persons. LAMBDA, since it is handcrafted makes sure that none of the videos are offensive in nature, UltraLAMBDA being sourced from the internet is noisier. While the videos themselves originate from brands, content of some brands may seem offensive to certain people.

We used Llama, GMHRA, ViT, EVA-CLIP, and Qformer models in accordance with their licenses to train Henry.

1. Create Super Lambda Dataset [OCR + ASR + Colors + SubsetEmotion + SubsetEmotion]
2. Train Vicuna1.5 Memorability model
3. Add PseudoLabels for Super Lambda

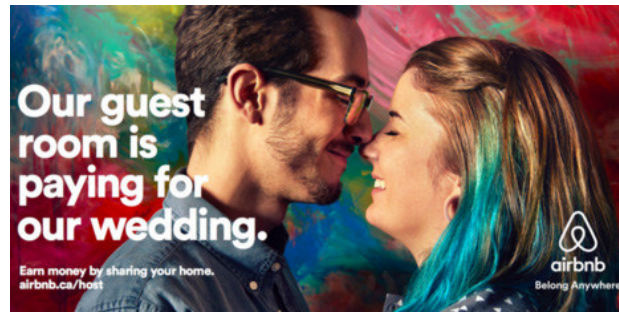


Figure 5: Airbnb advertisement showing the visual concepts of two adults, and the text “Our guest room is paying for our wedding”. “World knowledge” captured by LLMs helps identify the two adults as partners, and helps relate the text with the two adults and the Airbnb logo to infer what the ad is talking about.



Figure 6: The top three rows show the keyframes from videos in our dataset, LAMBDA, arranged from most to least memorable. The bottom two rows show brands arranged from the most memorable brands to the least.

4. RLHF for complete ads
5. denoising + on-the-fly backtranslation
6. self curation (scoring)

E1 Original Eng Sentence F1 Translated Eng Sentence

E1 -> F1 (inference only)

F1 -> E1* (train)

noise(E1) + F1 -> E1* (train) | on-the-fly backtranslation