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

arXiv:2309.00378v3 [cs.CL] 16 Feb 2024

Abstract

Marketers spend billions of dollars on advertisements, but to what end? At purchase time, 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. All previous memorability studies have been conducted on short-term recall on specific content types like object and action videos. On the other hand, the advertising industry only cares about long-term memorability, and ads are almost always highly multimodal. Therefore, we release the first 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, *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 novel model, Henry, to predict the memorability of a content which achieves state-of-the-art performance across *all* prominent literature memorability datasets. Henry 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 the LLM to generate more memorable ads. We show that the generated advertisements have 44% higher memorability scores than the original ads. Further, we release a large-scale ad dataset, UltraLAMBDA, consisting of 5 million ads with their automatically-assigned memorability scores.

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 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 (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.

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**

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. 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.

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.

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 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 (Atalay et al., 2023), emotional content (Putrevu et al., 2004; Mai and Schoeller, 2009), repetition (Schmidt and Eisend, 2015), spot length (Newstead and Romaniuk, 2010; Varan et al., 2020), the position of brand logo and imagery (Newstead and Romaniuk, 2010), sound presence (Bellman et al., 2021), and 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 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

---

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

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] (Fig. 2), incorporating world-knowledge from large language models (Llama (Touvron et al., 2023)), visual knowledge from vision encoder (EVA-CLIP (Sun et al., 2023)) 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 (Li et al., 2023) 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 (Khandelwal et al., 2024; Zheng et al., 2023). 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. Fine-tuning LLama for two iterations on this automatically constructed dataset yields an improvement of 44% in memorable ad generation.

Our main contributions are summarized as:

- 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).

- 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.

- 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 data exists.

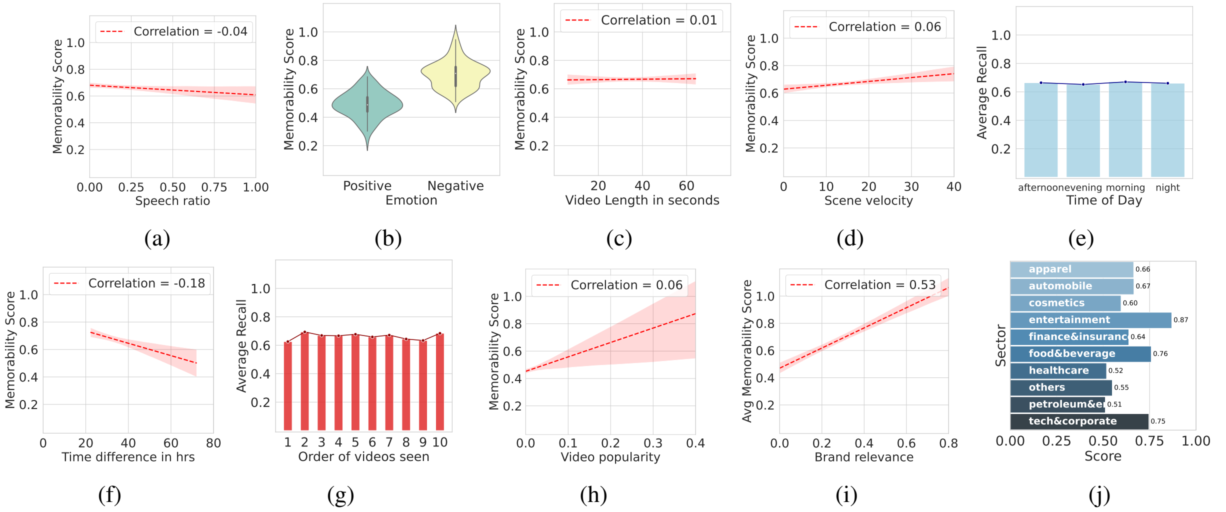
- We conduct an extensive set of experiments on memorability prediction, showing the effects of LTM on STM modeling and vice-versa, and the effects of changing world-knowledge with time, scene understanding, brand knowledge, and speech on memorability modeling (§3.3).

2 LAMBDA Protocol, Study & Insights

We first give an overview of LAMBDA data collection process and the annotation protocol. We also

---

[2] We 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.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

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.

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 (Newman et al., 2020; Cohendet et al., 2019), 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 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 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 are asked to describe what they remember about the ads of the recognized brands[3]. 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 the following factors: **brand factors** (*viz.,* brand popularity, industry), **content factors** (*viz.,* video emotion, scene velocity, length,

---

[3] The complete questionnaire for participant one is given in Appendix:§10.1.

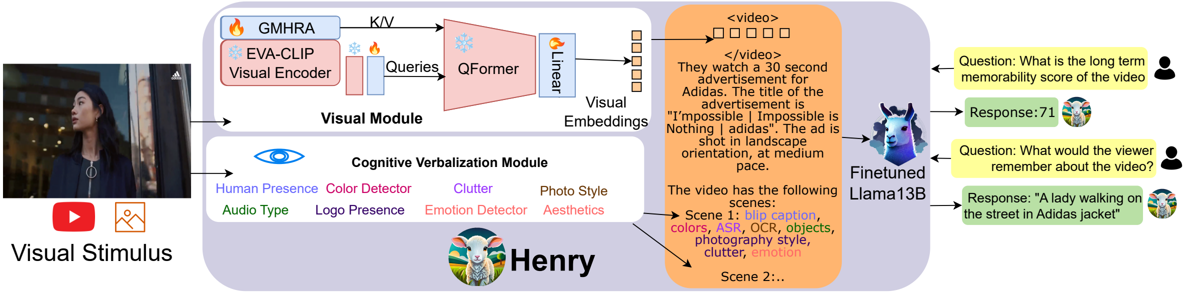

GMHRA
K/V
EVA-CLIP Visual Encoder
Queries
QFormer
Linear
Visual Module
Visual Embeddings
Cognitive Verbalization Module
Human Presence
Color Detector
Clutter
Photo Style
Audio Type
Logo Presence
Emotion Detector
Aesthetics
Henry
Visual Stimulus
<video>
</video>
They watch a 30 second advertisement for Adidas. The title of the advertisement is "I’mpossible | Impossible is Nothing | adidas". The ad is shot in landscape orientation, at medium pace.
The video has the following scenes:
Scene 1: blip caption, colors, ASR, OCR, objects, photography style, clutter, emotion
Scene 2:..
Finetuned Llama13B
Question: What is the long term memorability score of the video
Response:71
Question: What would the viewer remember about the video?
Response: "A lady walking on the street in Adidas jacket"

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.

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 (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, 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). Further, it may be expected that memorability decays as time passes. Comparing the 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. 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. 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, (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 (Kumar et al., 2023). World knowledge cap-

---

[4] see Appendix Fig. 10) 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 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: 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)

tured 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 (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 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 (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: To give visual knowledge to Henry, we use EVA-CLIP visual embedder (Sun et al., 2023). We find that Global Multi-Head Relation Aggregator (GMHRA) (Li et al., 2022) helps aggregate the ViT’s information better across the time dimension. Next, to effectively leverage the LLM’s rich language representations, we use Q-Former (Li et al., 2023) 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). The cognitive and visual features are given in Table 4 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 (Table 6).

3.2 Two-Stage Training

We do two-stage training 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 (see Listing 14). 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) (Ouyang et al., 2022) where we provide GPT

---

[5] Computing infrastructure used to conduct the experiments along with hyperparameters are given in Appendix:§12.1. All experiments are conducted with three random seeds and averages are reported.

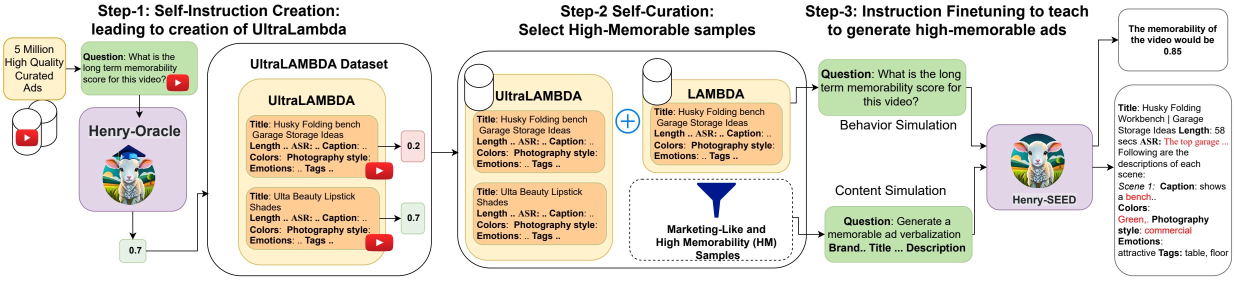

Step-1: Self-Instruction Creation:
leading to creation of UltraLambda
5 Million High Quality Curated Ads
Question: What is the long term memorability score for this video?
Henry-Oracle
0.7
UltraLAMBDA Dataset
UltraLAMBDA
Title: Husky Folding bench Garage Storage Ideas
Length .. ASR: .. Caption: ..
Colors: Photography style:
Emotions: .. Tags ..
0.2
Title: Ulta Beauty Lipstick Shades
Length .. ASR: .. Caption: ..
Colors: Photography style:
Emotions: .. Tags ..
0.7
Step-2 Self-Curation:
Select High-Memorable samples
UltraLAMBDA
Title: Husky Folding bench Garage Storage Ideas
Length .. ASR: .. Caption: ..
Colors: Photography style:
Emotions: .. Tags ..
Title: Ulta Beauty Lipstick Shades
Length .. ASR: .. Caption: ..
Colors: Photography style:
Emotions: .. Tags ..
LAMBDA
Title: Husky Folding bench Garage Storage Ideas
Length .. ASR: .. Caption: ..
Colors: Photography style:
Emotions: .. Tags ..
Marketing-Like and High Memorability (HM) Samples
Step-3: Instruction Finetuning to teach
to generate high-memorable ads
Question: What is the long term memorability score for this video?
Behavior Simulation
Content Simulation
Question: Generate a memorable ad verbalization
Brand.. Title ... Description
Henry-SEED
The memorability of the video would be 0.85
Title: Husky Folding Workbench | Garage Storage Ideas Length: 58 secs ASR: The top garage ... Following are the descriptions of each scene:
Scene 1: Caption: shows a bench..
Colors: Green,. Photography style: commercial
Emotions: attractive Tags: table, floor

Figure 3: Overview of our SEED method for memorable ad generation. Our self-alignment consists of three steps: (i) **Self-instruction creation**: We first collect 5 million high-quality ads from YouTube, Facebook, and other mediums. Henry-Oracle (trained on the complete train+test sets of LAMBDA) is then used to rate this curated set in an LLM-as-a-Judge fashion. (ii) **Self-curation**: We select marketing-like and high-memorability samples from the UltraLAMBDA and LAMBDA datasets. (iii) Instruction fine-tuning: Henry-SEED is trained on the self-curated set using two tasks: Behavior Simulation and Content Simulation.

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 (Hagen and Espeseth, 2023). 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. Tables 2 and Table 5 (in the appendix) show the data ablations. We see that combining datasets actually worsens the performance across all the datasets except the SUN dataset. Further, we find that in zero-shot settings, STM helps in predicting LTM relatively much better than vice versa. This corroborates with the studies in psychology which 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. Cognitive features that were found important in our study also improve prediction performance. Low-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 further propose a new task of memorable ad generation. Given the input as brand name, a brief campaign description, and the required ad duration, the task is to generate a memorable ad consisting of scene descriptions, characters, and dialogues. Most of the work in memorability has been about *how much* a content is memorable; there is little work to generate memorable content (Danescu-Niculescu-Mizil et al., 2012; Khosla et al., 2013; Siarohin et al., 2017; Goetschalckx et al., 2019), and most of it is in generating natural images, and not multimodal advertising videos. However, there does not exist any large enough dataset that can be leveraged to learn to generate memorable ads. Therefore, we propose our Self rEwarding mEmorability moDeling (SEED) method to learn to generate memorable ads.

SEED method (Fig. 3): *Step-1 Self-Instruction creation:* We first collect a dataset consisting of 5 million ads sourcing the dataset from social media channels, including Facebook, Twitter, Snapchat, and YouTube. We collect the brand name, ad title, ad links, posted caption, date, and raw ad bytes (video and images). We then automatically label the collected ads with the cognitive features important for modeling memorability (Table 4). Next, following LLM-as-a-judge (Zheng et al., 2023) and other related works that use synthetic labels from expert LLMs (Zhou et al., 2023; Li et al., 2024; Khandelwal et al., 2024), using Henry-Oracle[6], we automatically label the collected ad content for memorability scores. We call this dataset *UltraLAMBDA*.

Step-2 Self-Curation: Since the ads are crawled from publicly available ads, using few shot Mistral-7B (Jiang et al., 2023), we further filter the ads such that they are primarily marketing in nature, with proper language, semantics, and product relevance (Listing 16). Following this, we select high-memorable ads with a memorability score of > 65. *Step-3 Instruction Finetuning:* Next, we teach Henry (as trained in §3) two skills simultaneously: behavior simulation (given ad, predict memorability; Listing 14) and content simulation (given brand name, ad-title and required duration, generate ad scenes and dialogues; Listing 15). We call the model trained via the SEED procedure as Henry-SEED (Fig. 3).

Evaluation: We evaluate the generated ads on three metrics: **memorability** as judged by Henry-Oracle,

---

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

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	$UltraLAMBDA + LAMBDA_{High}$	820k	+89	+31	+12	+44	7.44	85.6%	60.48%
Henry-SEED	13B	SEED	$UltraLAMBDA$	800k	+41	+18	+1	+20	7.34	74.7%	-
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, humans preferred Henry-SEED ads more than the original ads 3/5 times while GPT-4 generated ads are preferred 2/5 times. Further, we note that an increase in the amount of training data for Henry-SEED increases its performance across all metrics. Figs. 4-6 and Listings 1-10 contain some qualitative samples generated using Henry-SEED.

ad quality as judged by GPT-4, and **ad quality** as judged by humans[7]. First, the brand names and titles are sourced from the test split of the LAMBDA dataset, and then models like GPT-4 and Henry-SEED are asked to generate ads given the brand and title. We evaluate the generated story's memorability score using Henry-Oracle. The expectation is that the memorability of the story generated by the story generation model should not be significantly worse than the original memorability score for high-memorable samples (score>65) but should be better than the scores for low (score<44) and medium memorability samples. Using GPT-4 as judge, we test two ad-quality metrics: consistency and preference. Consistency measures how consistent is the generated story with itself (for instance, subsequent dialogues with each other) and the consistency of story with the given title and brand information (Listing 12). Preference measures how many times GPT-4 prefers the generated story compared to the original story (Listing 11). In human-evaluation, we ask human annotators to choose between the generated and the original ad stories without telling them which one is generated or real (§9.1). We conduct the human evaluation using 20 non-expert annotators and 3 experts. The expectation is that ad quality for synthetic ads should be similar to the original ads.

Results: Table 3 shows results of Henry-SEED compared with in-context-learning based GPT-3.5 and GPT-4, and various ablations of the SEED approach. It can be noted that Henry-SEED performs much better in improving memorability scores across all three bins (Low, Medium, and High). The closest model is Henry-SEED trained only on $LAMBDA_{High}$. It is noteworthy that while 15x larger models, GPT-4 and GPT-3.5 increase the memorability on Low rated memorability ads, but they decrease the memorability on High rated ads. Comparing the ad quality, we see that while GPT-4 prefers its own generated ads 91.3% times, Henry-SEED is a close second with 85.6% preference scores. We test the ads generated by top two models using human annotators asking them to choose between the original and the generated ad based on ad quality. Henry-SEED's ads are preferred 60% times over the original ads, 20% more than GPT-4's ads. Figs. 4-6 and Listings 1-10 show some randomly sampled ads generated by Henry-SEED and Sec. 9.2 contains some expert comments over the generated ads. We also run some ablation studies to find the impact of the amount of data (Fig. 7), and the impact of behavior simulation and content simulation tasks (Table 7) on ad quality and memorability. We find that performance increases as the amount of data increases.

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.

---

[7] It is noteworthy that humans can't judge the memorability of a content (Isola et al., 2013); therefore, we ask them to evaluate only the ad quality.

6 Limitations and Potential Risks

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 a 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 a majorly uniform population, the training will be scaled to more languages and population types in 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. In our opinion, the model does not pose any potential risk or harm besides the limitations mentioned here. We also conduct a review of the generated ads through experts and non-expert annotators. Both experts and non-expert annotators preferred Henry-SEED generated ads 3/5 times.

References

Adobe. 2024. Adobe Firefly. `https://www.adobe.com/products/firefly.html`. Accessed: February 9, 2024.

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.

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*.

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*.

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-with-mo` 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.

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

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

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. 2023. Mistral 7b.

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.

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

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-Héléne 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, Yannis 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. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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*.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2024. Self-alignment with instruction backtranslation. *The Journal of Machine Learning Research*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. 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*.

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

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*.

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 Rezatofighi, and Dacheng Tao. 2022. Gmflow: Learning optical flow via global matching.

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

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. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, and Tong Sun. 2023. Customization assistant for text-to-image generation. *arXiv preprint arXiv:2312.03045*.

Appendix

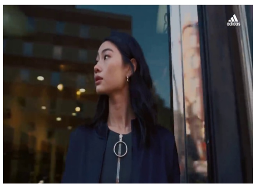
Image	Semantic Category	Verbalization	Semantic Category	Verbalization
	**OCR**	The text shown in the scene is "Adidas".	**Clutter**	The clutter in the scene is low.
	ASR	The audio in the scene is "To take hold of the world's spotlight overnight".	**Photo Style**	The photography style of the scene is 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 4: To augment the scene understanding of LLM, we verbalize video scenes and images using a diverse set of cognitive and 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., 2023), human presence (Liu et al., 2023), emotion (Singh et al., 2023), caption (Li et al., 2023), aesthetics (Ke et al., 2023), colors (Qin et al., 2020), object tags (Zhang et al., 2023), audio-type (Giannakopoulos, 2015), and logo presence (Zhang et al., 2023). Black colored text is the verbalization template, and red text indicates the model outputs.

7 Ablation Experiments For Henry

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. A few trends can be observed from the table: (i) STM generalizes better on LTM in zero-shot than vice versa (rows 1 and 2), (ii) Henry trained on either videos or images generalizes to both (rows 3 and 4), (iii) There is a significant performance loss in modeling memorability for brands not seen during training (row 5), (iv) Zero-shot generalization to Memento (video) and Memcat (image) is near to the current trained state of the art literature models on Memento (Dumont et al., 2023) and Memcat (Hagen and Espeseth, 2023) (rows 6 and 7).

	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 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 -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 -emotion	0.71	0.78	0.61	0.73	0.42	0.46
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
Henry -asr	-	-	-	-	-	0.46
Henry -asr-emotion	-	-	-	-	-	0.42

Table 6: Ablation across architectural choices. "-" denotes non-speech dataset. A few trends are visible from the table: (i) Despite having a vision branch, object tags and colors have a net positive impact on the overall performance (rows 2,3,4), (ii) For LTM, dropping cognitive features such as emotion, aesthetics, and clutter cause a larger performance drop than dropping visual features such as objects and colors. The trend is opposite for STM. This shows that visual features which are considered cognitvely lower level are more important for STM and higher level features are more important for LTM.

8 Ads Generated using Henry-SEED

Below are some qualitative examples of Ad generations from Henry. We provide visualizations for easier understanding, along with the raw generations (Listings 1-10). For the visualizations, we do the following:

1. Prompt Adobe Firefly (Adobe, 2024) with the exact scene descriptions provided by Henry
2. Add OCR from the generated verbalization on top of the generated images manually.
3. We segment the voice-over for each scene manually, since the current methodology provides the voice-over and scene description separately.

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

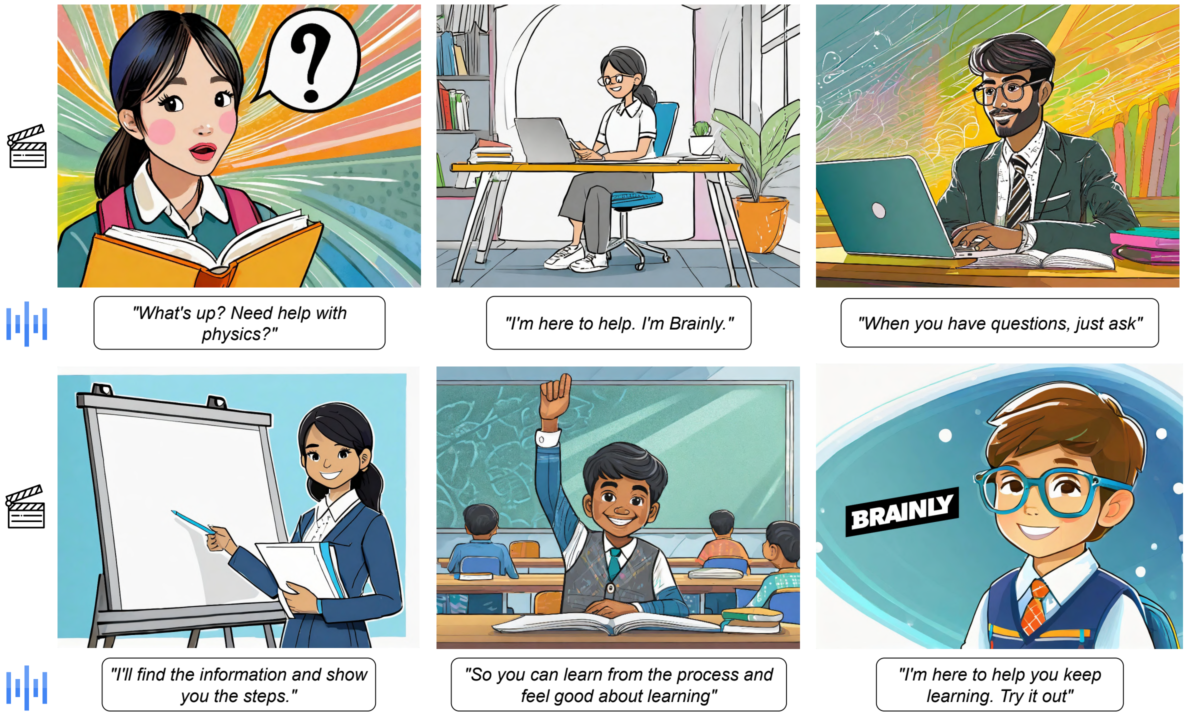
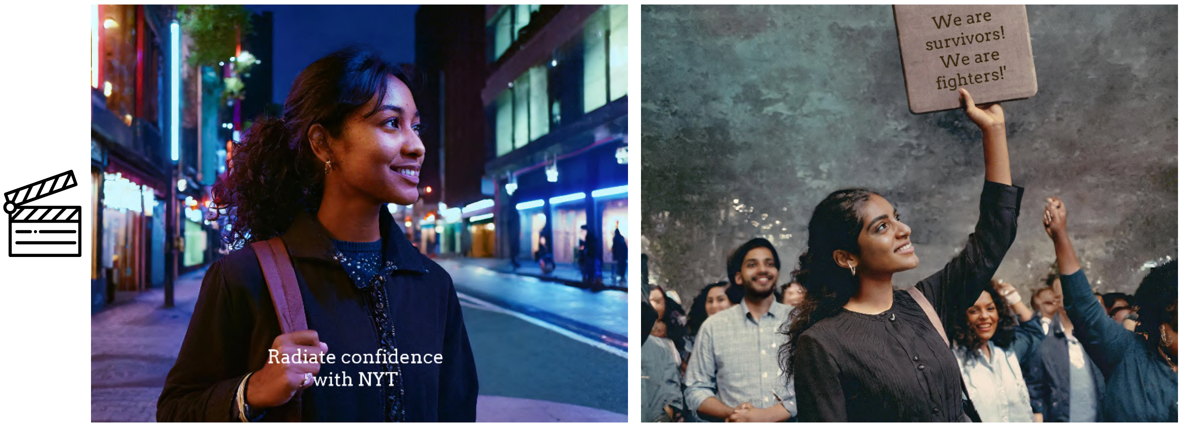
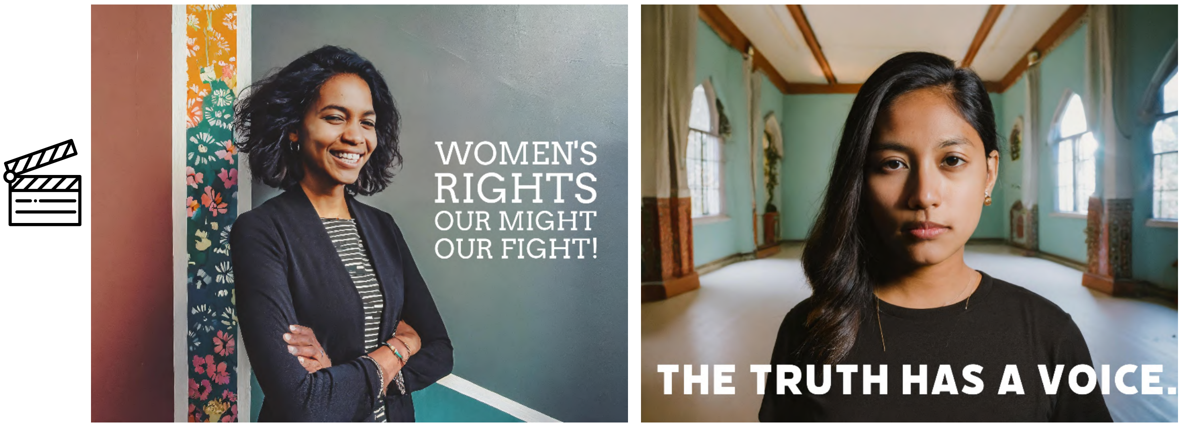

?
"What's up? Need help with physics?"
"I'm here to help. I'm Brainly."
"When you have questions, just ask"
BRAINLY
"I'll find the information and show you the steps."
"So you can learn from the process and feel good about learning"
"I'm here to help you keep learning. Try it out"

Figure 4: 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.



"The New York Times is shining a light on women's rights. For 150 years, The Times has held those in power accountable."

"From the fight for pay equality to the battle against harassment"



"The Times is telling the stories of women around the world. Join us in supporting our women."

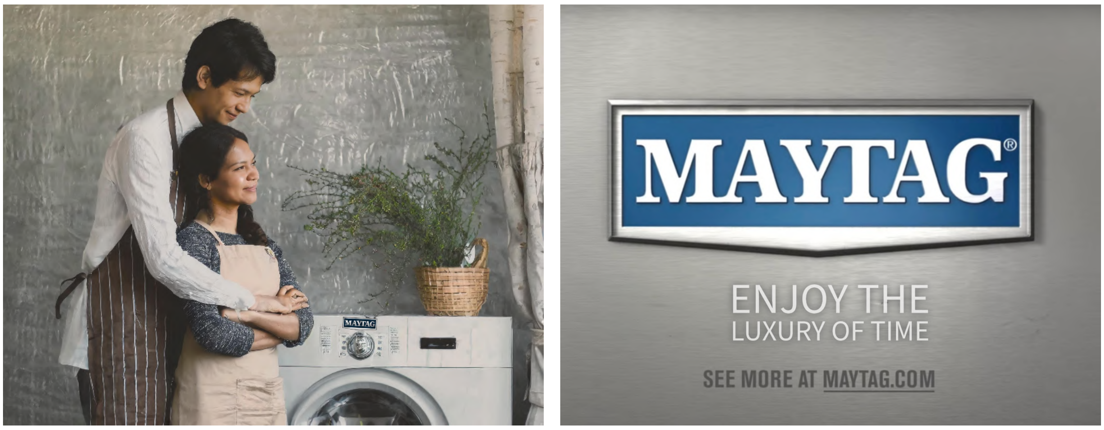
"Subscribe to The New York Times.
The truth has a voice."

Figure 5: Henry-SEED Prompt: *Generate the detailed description of a 50 second memorable advertisement titled "Shining a Light on Women's Rights | The Truth Has a Voice | The New York Times" for the brand The New York Times* Link to the original ad: `https://www.youtube.com/watch?v=bPblzhUzTeg` Original memorability score: 65. Memorability score of Generated Ad: 91.

"Maytag has taken the laundry room by storm with their overnight wash and dry."

"Maytag has taken the laundry room by storm with their overnight wash and dry."

"All without leaving the house. It's like having your own personal laundry assistant."



"Never fight over the laundry room again. Get a Maytag and enjoy the luxury of time."

"Maytag, Enjoy the luxury of time"

Figure 6: Henry-SEED Prompt: *Generate the detailed description of a 18 second memorable advertisement titled "Maytag Overnight Wash and Dry" for the brand Costco*. Link to the original ad: `https://www.youtube.com/watch?v=uT72lJhUUS0` Original memorability score: 76. Memorability score of the generated Ad: 83

Given below are some examples of advertisements generated by Henry-SEED. Both input and output are given in the Listings 1-10.

Listing 1: Advertisement generated by Henry-SEED for the Costco Ad. Original Ad: https://www.youtube.com/watch?v=uT72lJhUUS0

```
Input Title: "Maytag Overnight Wash and Dry" for the brand
      Costco


Generated Ad: "The ad is shot in landscape orientation, at a
      medium pace.

Scene 1: The scene shows a person standing in front of a
      washer and dryer
Emotions: comforting, comfortable
Tags: appliance, laundry, laundry room, man
Voiceover: "Maytag has taken the laundry room by storm with
      their overnight wash and dry."
Colors: Gray, Off_White, Silver, White, Black, Dark_Gray

Scene 2: The scene shows a person lying on a bed pressing a
      button on a cell phone
Emotions: gift, birthday, holiday
Tags: button, catch, smartphone, hand
Voiceover: "Imagine going to bed with dirty clothes and waking
       up to a fresh, clean load. All without leaving the house
      . It's like having your own personal laundry assistant."
Colors: Gray, Silver, Black, Dark_Gray, Dark_Brown

Scene 3: The scene shows a man and woman embracing while a
      washer and dryer are in the background
Emotions: love, happiness, togetherness
Tags: couple, hug, laundry, laundry room
Voiceover: "Never fight over the laundry room again."
Colors: Gray, Off_White, Silver, White,Black, Dark_Gray

Scene 4: The scene shows a logo of Maytag
Emotions: trust, confidence
Tags: logo, Maytag
Voiceover: "Get a Maytag and enjoy the luxury of time."
Colors: Off_White, Silver, White, Black, Dark_Gray, Gray
"
```

Listing 2: Advertisement generated by Henry-SEED for the Pruduential Financial Ad. Original Ad: https://www.youtube.com/watch?v=JPfi0P-0nwU

```
Input Title:  "Financial Professional Tips: Planning for Your
       Own Care" for the brand Prudential Financial

Generated Ad: "The ad is shot in landscape orientation, at a
      low pace.

Scene 1: The scene shows a woman looking thoughtfully at a
      laptop on her desk
Emotions: concern and contemplation
Tags: computer, laptop, office, woman
Colors: Gray, Dark_Blue

Scene 2: The scene shows the same woman now with a slight
      frown as she examines papers that appear to be medical
      bills
Emotions: worry and realization
Tags: bills, healthcare, paperwork, desk
Voiceover: "Long-Term Care is one of the most important issues
       in financial planning, but it's not always easy to talk
      about."
Colors: Gray, Dark_Blue

Scene 3: The scene shows the woman engaging in a conversation
      with a financial professional
Emotions: engagement and interest
Tags: consultation, financial advisor, discussion, office
Voiceover: "As a Financial Professional, I can help you
      evaluate your options and determine how much you may need
       to set aside each month."
Colors: Olive, Gray

Scene 4: The scene shows the woman in deep thought, looking
      out of a window while holding a pen and notepad,
      suggesting planning and reflection
Emotions: contemplation and seriousness
Tags: planning, reflection, window, notepad
Colors: Sky_Blue, Light_Gray

Scene 5: The scene transitions back to show the woman nodding
      in understanding as she listens to the financial
      professional, visibly gaining confidence
Emotions: comprehension and confidence
Tags: financial professional, office, learning, agreement
Voiceover: "Most people believe that they can rely on Medicare
       to pay for their long-term care expenses. But the truth
      is, Medicare typically only covers about 65%
Colors: Gray, Dark_Blue

Scene 6: The scene shows the woman now jotting down notes
      actively, with a look of determination on her face
Emotions: determination and proactivity
Tags: planning, notetaking, strategy, office
Voiceover: "So if you're a typical American, you'll need about
       $100,000 to cover the remaining costs of long-term care
      ."
Colors: Gray, Dark_Blue

Scene 7: The scene shifts to the woman shaking hands with the
      financial professional, symbolizing a plan set in motion
Emotions: satisfaction and trust
Tags: handshake, agreement, financial planning, commitment
Colors: Gray, Dark_Blue

Scene 8: The scene cuts to the woman now at home, relaxed and
      smiling, speaking on the phone, perhaps sharing her
      relief and positive steps taken with a loved one
Emotions: relief and happiness
Tags: home, communication, personal life, reassurance
Voiceover: "Let's talk about a plan for your own care."
Colors: Cream, Soft_Green

Scene 9: The final scene zooms out to a serene image of the
      woman walking in a park, symbolizing peace of mind about
      her future
Emotions: contentment and peace
Tags: park, walking, future, peace of mind"
Colors: Sky_Blue, Green
```

Listing 3: Advertisement generated by Henry-SEED for an Adidas Ad. Original Ad: https://www.youtube.com/watch?v=fujboh-W7Sk

```
Input Title:  "I'M POSSIBLE | IMPOSSIBLE IS NOTHING | Adidas"
      for the brand Adidas


Generated Ad: "The ad is shot in landscape orientation, at a
      medium pace.

Scene 1: The scene shows a man facing a rock climbing wall,
      poised to start his ascent
Emotions: anticipation and focus
Tags: athletic, climb, climber, rock
Voiceover: "What is possible is up to you."
Colors: Gray, Dark_Brown

Scene 2: The scene shows a man as he begins to climb, gripping
       the rock tightly
Emotions: determination and strain
Tags: athletic, climb, climber, hand
Voiceover: "Don't let anyone tell you otherwise."
Colors: Dark_Gray, Dark_Brown

Scene 3: The scene shows the man midway up the rock wall,
      reaching for a distant handhold
Emotions: perseverance and challenge
Tags: athletic, climb, climber, hand
Colors: Black, Dark_Gray

Scene 4: The scene captures a moment where the man slips but
      regains his grip on the rock face, muscles tensing
Emotions: tension and danger
Tags: athletic, climb, climber, hand
Colors: Dark_Gray, Dark_Brown

Scene 5: The scene shows the man pausing to chalk his hands,
      his breath visible in the cool air
Emotions: calm and preparation
Tags: athletic, chalk, climb, climber
Colors: Gray, Dark_Brown

Scene 6: The scene shows the man navigating a particularly
      difficult part of the climb, his movements precise and
      deliberate
Emotions: concentration and determination
Tags: athletic, climb, climber, hand
Colors: Black, Dark_Gray

Scene 7: The scene shows the climber's face, sweat beading on
      his forehead, reflecting the intensity of his effort
Emotions: grit and resolve
Tags: athletic, climb, climber, determination
Colors: Dark_Gray, Dark_Brown

Scene 8: The scene shows the man reaching for the final ledge,
       the clear sky behind him signaling the nearing end of
      his ascent
Emotions: anticipation and effort
Tags: athletic, climb, climber, hand
Voiceover: "Impossible is not an obstacle."
Colors: Dark_Gray, Gray, Dark_Brown, Blue
```

```
Scene 9: The scene shows the man pulling himself onto the
      summit, his body silhouetted against the sky
Emotions: triumph and exhilaration
Tags: athletic, climb, climber, hand
Voiceover: "It's a chance to prove them wrong."
Colors: Dark_Gray, Dark_Brown, Blue

Scene 10: The scene shows the man standing on top of the rock
      with his arms raised in victory, overlooking the
      landscape below
Emotions: achievement and elation
Tags: athletic, arms up, climb, climber"
Colors: Dark_Gray, Dark_Brown
```

Listing 4: Advertisement generated by Henry-SEED for an Amazon ad. Original Ad: `https://www.youtube.com/watch?v=rcFTqs-zSH4`

```
Input Title: "Alexa Auto, Listening to Audiobooks" for the
      brand Amazon

Generated Ad: "The ad is shot in landscape orientation, at a
      low pace.

Scene 1: The scene shows a car driving down a country road
      near a mountain house
Emotions: anticipation, freedom
Tags: blue, car, drive, house
Colors: White, Lavender, Dark_Blue,Turquoise, Black

Scene 2: The scene shows an image of a car approaching a large
       sign above the road, signaling the start of the highway
Emotions: curiosity, readiness
Tags: car, drive, road sign, highway
Voiceover: "You just pulled onto the highway"
Colors: White, Lavender,Turquoise, Gray

Scene 3: The scene shows an in-car view with a driver's hands
      on the steering wheel, a smartphone mounted on the
      dashboard displaying the Amazon Music app
Emotions: concentration, engagement
Tags: car interior, hands, smartphone, technology
Voiceover: " but you can't remember the name of that band that
      's been stuck in your head all morning."
Colors: White, Dark_Blue,Black, Gray

Scene 4: The scene shows the driver speaking to Alexa through
      the car's infotainment system, the display lighting up in
       response
Emotions: interaction, ease
Tags: car interior, voice command, infotainment system
Voiceover: "With Alexa and Amazon Music, you don't have to
      remember.Just ask, "Alexa, play the Beatles."
Colors: White, Lavender, Dark_Blue, Black

Scene 5: The scene shows the car passing through a landscape
      with rolling hills and the sun beginning to set in the
      background
Emotions: satisfaction, serenity
Tags: car, drive, hills, sunset
Colors: White, Lavender, Turquoise, Dark_Blue

Scene 6: The final scene shows the vehicle driving into the
      horizon, as the iconic music of the Beatles fills the car
      's interior
Emotions: fulfillment, joy
Tags: car, drive, horizon, music"
Colors: White, Lavender, Orange, Dark_Blue
```

Listing 5: Advertisement generated by Henry-SEED for an Gatorade ad. Original Ad: `https://www.youtube.com/watch?v=BKPQkjRF4yY`

```
Input Title: "Gatorade | Make Your Own Footsteps with Suni
      Lee" for the brand Gatorade

Generated Ad: "The ad is shot in landscape orientation, at a
      medium pace.

Scene 1: The scene shows Sunisa Manado standing in front of
      her house
Emotions: determination, brave, confident, persistence,
      courage
Tags: athletic, face, girl, muscle
Voiceover: "I'm Sunisa Manado, but you can call me Sunisa. And
       I'm going to show you how I stay motivated. Let's go!
      This is my neighborhood. I'm going to take you through my
       favorite spots. There's my house right there."
Colors: Dark_Brown, Black, Brown, Tan

Scene 2: The scene shows Sunisa Manado performing a handstand
      in the park
Emotions: achievement, determination, persistence, commitment,
       success
Tags: balance, gymnast, handstand, girl
Voiceover: "And this is the park where I get so much done.
      This is the park where I train."
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 3: The scene shows Sunisa Manado doing a flip on the
      balance beam
Emotions: brave, courage, determination, persistence,
      inspiration
Tags: gymnast, flip, beam, girl
Voiceover: "Being an athlete takes a lot of hard work and
      determination."
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 4: The scene shows Sunisa Manado in a powerful pose in
      her pink sports bra and leotard
Emotions: determination, brave, courage, persistence,
      inspiration
Tags: athletic, face, girl, gymnast
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 5: The scene shows Sunisa Manado lifting herself on the
      parallel bars
Emotions: achievement, persistence, determination, courage,
      commitment
Tags: gymnast, lift, bars, girl
Voiceover: "And being an athlete also means that you have to
      have good nutrition."
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 6: The scene shows Sunisa Manado practicing her routine
      on the floor of the park
Emotions: determination, persistence, inspiration, commitment,
       discipline
Tags: floor, gymnast, routine, girl
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 7: The scene shows Sunisa Manado spinning on the balance
      beam
Emotions: success, brave, determination, precision, focus
Tags: beam, spin, gymnast, girl
Colors: Dark_Brown, Dark_Blue, Purple, Gray

Scene 8: The scene shows Sunisa Manado holding a bottle of
      Gatorade, rehydrating after her practice
Emotions: refreshed, healthy, satisfaction, wellbeing,
      recovery
Tags: athlete, drink, Gatorade, health
Voiceover: "That's why I choose to fuel up with Gatorade. You
      need to have that energy, especially during my toughest
      training."
Colors: Black, Dark_Green, Dark_Gray, Gray

Scene 9: The scene shows Sunisa Manado smiling towards the
      camera with a sense of pride and accomplishment
Emotions: joy, confidence, pride, optimism, accomplishment
Tags: face, smile, girl, gymnast
Voiceover: "Gatorade helps replenish what I've lost and helps
      keep me going."
Colors: Black, Dark_Green, Dark_Gray, Gray

Scene 10: The scene shows Sunisa Manado walking away from the
      park, as the sun sets, symbolizing the end of a day's
      hard work
Emotions: contemplation, fulfillment, anticipation, resilience
      , hope
Tags: sunset, walk, reflection, girl
Voiceover: "I'm going to make my own footsteps. Follow your
      dreams, but don't forget to chase your greatness."
Colors: Black, Dark_Green, Dark_Gray, Gray"
```

Listing 6: Advertisement generated by Henry-SEED for a Rolex ad. Original Ad: `https://www.youtube.com/watch?v=vA-Lxmh1MdQ`

```
Input Title: "Rolesor A Meeting of Metals" for the brand Rolex

Generated Ad: "The ad is shot in landscape orientation, at a
      low pace."

Scene 1: The scene shows a close up of a watch that is on a
      wrist
Emotions: anticipation, luxury, elegance
Tags: dial, gold, watch
Colors: Dark_Brown, Brown, Black, Khaki

Scene 2: The scene shows a gold and silver watch sitting on a
      table, reflecting soft light
Emotions: sophistication, status, success
Tags: dial, gold, watch
Colors: Brown, Dark_Brown, Khaki, Black

Scene 3: The scene shows a person's hands elegantly wrapping
      the gold and silver watch in delicate gift paper
Emotions: care, generosity, surprise
Tags: wrapping, gift, luxury
```

```
Colors: Brown, Khaki, Gold, White

Scene 4: The scene shows a recipient's joyous reaction as they
       receive the watch as a gift at a celebratory gathering
Emotions: happiness, gratitude, celebration
Tags: emotion, celebration, gift
Colors: Gold, Silver, Cream, Warm hues, Soft lighting

Scene 5: The scene shows the watch being worn at an elegant
       event, the gold and silver glinting under ambient
       lighting
Emotions: pride, accomplishment, refinement
Tags: elegance, luxury, watch
Colors: Gold, Silver, Black, Soft lighting, Dark hues
```

Listing 7: Advertisement generated by Henry-SEED for an NVIDIA ad. Original Ad: https://www.youtube.com/watch?v=HnaOhOJ6AcM

```
Input Title: "Pac Man 256 on NVIDIA SHIELD" for the brand
       Nvidia

Generated Ad: "The ad is shot in landscape orientation, at a
       medium pace. "

Scene 1: The scene shows a black and white logo.
Emotions: danger, tough
Tags: brand, logo
The text shown in the scene is 'SHIELD'
Colors: Black, White, Gray

Scene 2: The scene shows a black screen with a red light on it
Emotions: danger, tough
Tags: brand, logo, text
The text shown in the scene is 'SHIELD'
Colors: Black, White, Red

Scene 3: The scene shows a computer screen showing a game of
       Pac-Man
Emotions: nostalgia, fun
Tags: arcade, arcade machine, game, Pac-Man
The text shown in the scene is 'HIGH SCORE'
Colors: Cyan, Yellow, Gray, Dark Green, Red

Scene 4: The scene shows a video game character navigating a
       maze filled with obstacles
Emotions: excitement, challenge
Tags: arcade, character, game, maze
The text shown in the scene is 'LEVEL UP'
Colors: Cyan, Red, Pink, Orange, Gray

Scene 5: The scene shows a video game with a racing theme,
       cars zooming on the track
Emotions: thrill, speed
Tags: racing, cars, game, screenshot
The text shown in the scene is 'NEW RECORD'
Colors: Red, Yellow, Blue, Gray, Black

Scene 6: The scene shows a strategy video game with a detailed
       world map
Emotions: focus, strategy
Tags: strategy game, world map, empire, video game
The text shown in the scene is 'EXPAND YOUR EMPIRE'
Colors: Green, Brown, Blue, White, Tan

Scene 7: The scene shows a futuristic virtual reality game
       with immersive graphics
Emotions: amazement, innovation
Tags: VR, virtual reality, future, game
The text shown in the scene is 'ENTER NEW WORLDS'
Colors: Neon Blue, Neon Green, Black, Dark Blue

Scene 8: The scene shows a family gathered around playing a
       multiplayer video game, laughing and enjoying
Emotions: joy, togetherness
Tags: family, multiplayer, game, living room
The text shown in the scene is 'PLAY TOGETHER'
Colors: Multi-colored, Warm Hues

Scene 9: The scene transitions to a player wearing headphones,
       deeply concentrated in an esports competition
Emotions: determination, competition
Tags: esports, competition, gamer, video game
The text shown in the scene is 'RISE TO THE CHALLENGE'
Colors: Black, Neon, Dark, Spotlights

Scene 10: The scene closes with the NVIDIA SHIELD logo
       emerging brightly on the screen
Emotions: empowerment, assurance
Tags: brand, logo, closing
The text shown in the scene is 'NVIDIA. SHIELD - Game Your Way
       '
Colors: Red, White, Black
```

Listing 8: Advertisement generated by Henry-SEED for a Zillow ad. Original Ad: https://www.youtube.com/watch?v=P6gydrvTkaA

```
Input Title: "Find It, Tour It and Own It With Zillow (:06)"
       for the brand Zillow

Generated Ad: "The ad is shot in landscape orientation, at a
       medium pace.

Scene 1: The scene shows a person holding a cell phone with
       their hand.
Emotions: achievement, ambition
Tags: catch, smartphone, hand, person, phone, screen, text
       message, woman
Colors: Gray, Dark_Blue, Dark_Brown, Dark_Pink, Off_White

Scene 2: The scene shows a woman in a sweater looking at a
       cell phone.
Emotions: happy, smile
Tags: smartphone, earphone, girl, person, smile, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue

Scene 3: The scene shows a woman looking at a phone with a map
       on it.
Emotions: confident, smiling
Tags: catch, smartphone, earphone, girl, person, phone, stand,
       talk, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue

Scene 4: The scene shows a woman looking at a cell phone with
       a real estate app on it.
Emotions: togetherness, happiness
Tags: smartphone, earphone, girl, person, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue

Scene 5: The scene shows a woman looking at a cell phone with
       a real estate app displayed.
Emotions: happy, smiling
Tags: smartphone, earphone, girl, person, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue

Scene 6: The scene shows a woman using a cell phone to speak
       with an agent.
Emotions: confident, happy
Tags: smartphone, earphone, girl, person, talk, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue
Voiceover: "Getting the perfect home is a journey, so we help
       you find it, tour it, and own it."

Scene 7: The scene shows a woman completing a transaction on a
       cell phone.
Emotions: achievement, satisfied
Tags: smartphone, earphone, girl, person, woman
Colors: Gray, Dark_Brown, Dark_Pink, Dark_Blue"
Voiceover:  "For moving made simple, there's no place like
       Zillow"
```

Listing 9: Advertisement generated by Henry-SEED for a Kroger ad. Original Ad: https://www.youtube.com/watch?v=SqwqI01q3fA

```
Input Title: "How to Make Taco Seasoning | Kroger Recipes |
       Kroger" for the brand Kroger

Generated Ad: "The ad is shot in landscape orientation, at a
       low pace.

Scene 1: The scene shows a person pouring chipotle adobo sauce
       into a glass jar.
Emotions: care, comfort
Tags: bottle, can, container, hand, food, person, jar, liquid,
       pepper, pour, red, sauce, tomato sauce, tin, tray, woman
Colors: Black, Gray, Dark_Brown, Maroon, Dark_Red
Voiceover: "The audio in the ad says Taco seasoning is one of
       those spices that everyone loves on their food. It's so
       delicious and it's so easy to make."

Scene 2: The scene shows a woman in a red sweater adding
       ground cumin to the mix.
Emotions: anticipation, focus
Tags: blender, container, food, hand, person, ingredient,
       measuring cup, spice, woman
Voiceover: "All you have to do is get your chopstick and add
       some of the garlic and some of the onion and some of the
       cumin"
Colors: Black, Gray, Dark_Brown, Maroon, Dark_Red

Scene 3: The scene shows the addition of chili powder to the
       seasoning mix.
Emotions: precision, satisfaction
Tags: blender, container, food, food processor, hand, person,
       ingredient, mixture, spice, woman
Voiceover: " and some of the dried chilies and some of the
       oregano and some of the salt"
Colors: Black, Gray, Dark_Brown, Dark_Pink, Maroon
```

```
Scene 4: The scene shows the woman sprinkling crushed red
      pepper into the mix.
Emotions: excitement, zeal
Tags: blender, container, food, food processor, hand, person,
      ingredient, measuring cup, mixture, spice, woman
Voiceover: "and some of the paprika and some of the pepper and
       just give it a little bit of a squeeze"
Colors: Black, Gray, Dark_Brown, Maroon, Dark_Red

Scene 5: The scene shows the woman measuring and adding ground
       oregano to the seasoning mix.
Emotions: happiness, creativity
Tags: blender, container, food, food processor, hand, person,
      ingredient, measuring cup, mixture, spice, woman
Colors: Black, Gray, Dark_Brown, Maroon, Dark_Red

Scene 6: The scene shows the final steps of making the taco
      seasoning mix, with the woman adding garlic powder and
      onion powder.
Emotions: accomplishment, pride
Tags: blender, container, food, food processor, hand, person,
      ingredient, mixture, spice, woman
Voiceover: "You're going to taste it, so don't be afraid to
      add more or less of any of these ingredients, depending
      on how spicy or how flavorful you want your taco
      seasoning to be."
Colors: Black, Gray, Dark_Brown, Maroon, Dark_Red"
```

Listing 10: Advertisement generated by Henry-SEED for a Publix ad. Original Ad: https://www.youtube.com/watch?v=HH0VWvs-cFg

```
Input Title: "GreenWise Ice Cream Scoop from Publix" for the
      brand Publix Supermarkets

Generated Ad: "The ad is shot in landscape orientation, at a
      low pace.

Scene 1: The scene shows a woman standing in front of a
      refrigerator.
Emotions: anticipation, satisfaction
Tags: beverage, bottle, catch, doorway, drink, fill, girl,
      home appliance, juice, laugh, pour, refrigerator, shirt,
      stand, woman
Voiceover: "Greenwise, the better energy choice from Publix."
Colors: Black, Gray, Dark_Gray, Dark_Brown, Silver

Scene 2: The scene shows a woman holding a roll of Greenwise
      ice cream in her hand.
Emotions: curiosity, interest
Tags: bottle, catch, hand, woman
Voiceover: " All the deliciousness of ice cream, a fraction of
       the calories. And now it comes in a roll."
Colors: Black, Gray, Dark_Gray, Dark_Brown

Scene 3: The scene shows the woman as she easily scoops the
      ice cream onto a plate, displaying the convenience of the
      new roll format.
Emotions: ease, delight
Tags: plate, roll, scoop, serve, woman
Voiceover: " So you can easily scoop and serve exactly what
      you need."
Colors: Black, Gray, Dark_Gray, Dark_Brown

Scene 4: The scene shows the woman enjoying a bite of the
      Greenwise ice cream, her expression reflecting the
      product's deliciousness.
Emotions: enjoyment, satisfaction
Tags: bite, enjoyment, ice cream, satisfaction, taste, woman
Voiceover: "It's a win-win. Learn more at publix.com."
Colors: Black, Gray, Dark_Gray, Dark_Brown"
```

9 More Results for Henry-SEED

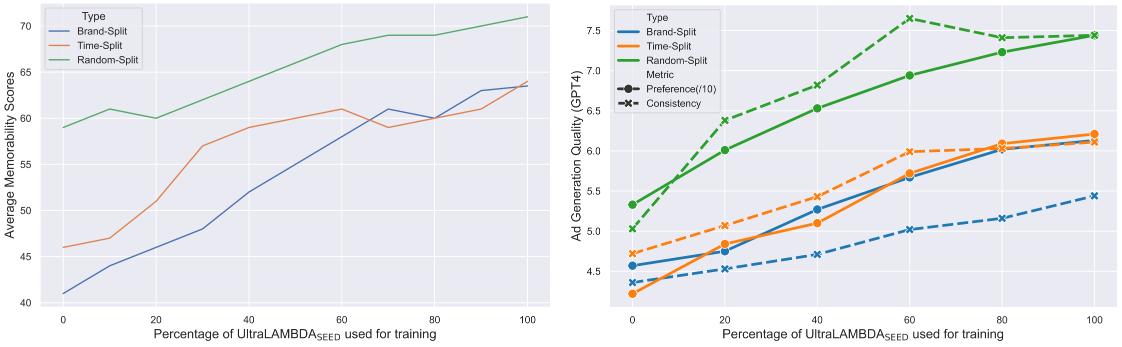
9.1 Questionnaire to Gather Human Preferences over Generated Ads

Below is the web-based form used to annotate the human preferences between the generated and original ads. Participants for this task were working professionals in the software, marketing, advertising, and creative industries. Participation was voluntary, and participants were invited to judge the efficacy of generated advertisements. Participants had a general interest in the creative and advertising industries and generative technologies; therefore, they were not interested in getting paid but rather in seeing and trying out the generative technology stack. We have a roughly 65-35 distribution of males to females with the age range between 22-50.

Task	LAMBDA (ρ)	Δ Memorability
BS-only	0.541	-
CS-only	-	+28.41
BS+CS	0.547	+30.66

Table 7: Ablation on modeling behaviour simulation (BS) or memorability prediction and Content Simulation (CS) on memorable ad generation together. For memorability prediction, we again show the Spearman rank correlation on the test set similar to 2, for generation we measure the change in memorability according to Henry Oracle similar to 3. We ovserve that mixing the two tasks together increases the performance across both tasks.

```
Instructions:

Shown next are 10 pairs of advertisements. Determine which ad
      within each pair is more effective based on the title,
      brand, and scene-by-scene descriptions provided. You will
      also be expected to provide reasons for your choice
      wherever asked.


Question 1
Choose the advertisement you find more effective. Also provide
       reasons for your choice.

Title: Bike to Work Day at NVIDIA
Brand: Nvidia
Nvidia is a technology company focusing on graphics processing
       units (GPUs) for gaming, professional visualization,
      data centers, and automotive markets, driving innovation
      in visual computing.

Advertisement A:

The ad is shot in landscape orientation, at a medium pace. The
       audio in the ad is silent.
Scene 1: The scene shows the camera takes a photo from the
      inside of the person on the bicycle
Colors: White, Dark_Pink, Olive, Gray, Pink, Dark_Brown
Emotions: danger, dangerous, warning
Tags: attach, bicycle, catch, smartphone

Scene 2: The scene shows the person riding a bicycle down the
      road
Colors: White, Dark_Gray, Mud_Green, Olive, Gray
Emotions: danger, quiet
Tags: bicycle, path, grass, motorbike
The text shown in the scene is 'NVIDIA'

Scene 3: The scene shows a man on a bike taking a ride
Colors: Off_White, Dark_Gray, Silver, Black, Gray
Emotions: danger, exciting, fun
Tags: bicycle, biker, bridge, hand
The text shown in the scene is 'DVIDIA'

Scene 4: The scene shows a bike rider going under a bridge
      under a road
Colors: Dark_Gray, Silver, Light_Green, Green, Olive, Gray,
      Bright_Green
Emotions: danger, dangerous, funny
Tags: bridge, car, curve, highway
The text shown in the scene is 'NVIDIA'

Scene 5: The scene shows a man riding a bicycle down a tree
      lined street
Colors: White, Dark_Gray, Mud_Green, Dark_Pink, Olive, Black,
      Gray
Emotions: thrill, adventure, romantic
Tags: bicycle, biker, hand, person
The text shown in the scene is 'NVIDIA'

Scene 6: The scene shows a man riding on a bicycle down the
      street
Colors: Emerald, Dark_Gray, Silver, Light_Green, Olive, Gray
Emotions: funky, enjoyable
Tags: bicycle, hand, person, man
The text shown in the scene is 'NVIDIA'

Scene 7: The scene shows a closeup of someone riding a bicycle
       down a road
Colors: White, Dark_Gray, Silver, Dark_Pink, Olive, Gray
```



(a) (b)

Figure 7: Graphs showing the importance of the amount of synthetic data on (i) Ad memorability score and (ii) Ad quality for the generated ads. As we can see from the graphs, both the ad memorability and quality increase with the increase in the amount of synthetic data.

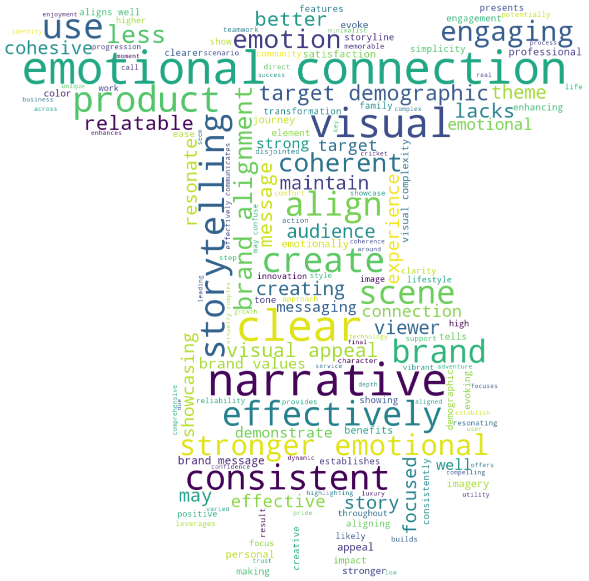


Figure 8: Word Cloud (resembling Henry) for the GPT-4 reasoning on the 75/88 generations where it rates Henry-SEED Generated Ads to be better than the Original.

```
Emotions: danger, majestic
Tags: bicycle, bicycle helmet, biker, hand
The text shown in the scene is 'NVIDIA'

Scene 8: The scene shows a person is riding a bike on the side
        of the road
Colors: White, Dark_Gray, Mud_Green, Olive, Gray, Lavender
Emotions: enjoy, enjoyment
Tags: car, person, man, motorcycle
The text shown in the scene is 'NVIDIA'

Scene 9: The scene shows someone riding a bike in front of a
        small city
Colors: White, Dark_Gray, Olive, Black, Gray
Emotions: funky
Tags: bicycle, biker, bin, car
The text shown in the scene is 'NVIDIA'

Scene 10: The scene shows a cyclist riding his bike on a
        gravel road
Colors: White, Brown, Mud_Green, Olive, Gray, Dark_Brown, Cyan
Emotions: recreational, adventure
Tags: bicycle, biker, hand, person

Advertisement B:

The ad is shot in landscape orientation, at a low pace. The
        audio in the ad is silent.

Scene 1: The scene shows a man wearing a hard hat holding a
        bike helmet
Colors: Dark_Gray, Brown, Mud_Green, Cream, Olive, Black,
        Dark_Brown
Emotions: protective, protective
Tags: building, construction worker, hat, jumpsuit

Scene 2: The scene shows a man riding a bike on a path near a
        creek
Colors: Emerald, Dark_Gray, Mud_Green, Olive, Black,
        Dark_Brown
Emotions: recreational, relaxation
Tags: bicycle, bicycle helmet, biker, path

Scene 3: The scene shows a man holding a bike up while
        standing in front of a building
Colors: Dark_Gray, Brown, Mud_Green, Cream, Olive, Black,
        Dark_Brown
Emotions: pride, achievement
Tags: building, professional, hat, bicyclist

Scene 4: The scene shows a man riding a bike down a street
        with trees lining the road
Colors: Brown, Cream, Green, Olive, Dark_Brown
Emotions: cheery, freedom
Tags: bicycle, bicycle helmet, biker, man

Scene 5: The scene shows a man riding a bike down a street in
        front of a house
Colors: Dark_Gray, Mud_Green, Olive, Black, Dark_Brown
Emotions: cheery
Tags: bicycle, bicycle helmet, biker, car

Scene 6: The scene shows a closeup of the man's face as he
        adjusts his bike helmet, showcasing determination
Colors: Cream, Olive, Black, Gray, Dark_Brown
Emotions: determined, prepared
Tags: man, helmet, focus, detail

Scene 7: The scene shows the man holding his bike next to
        other cyclists at a traffic light, promoting community
```

```
      and camaraderie
Colors: Mud_Green, Cream, Olive, Dark_Brown
Emotions: community, anticipation
Tags: cyclists, traffic light, group, waiting

Scene 8: The scene shows the man arriving at work, parking his
      bike in a bike rack
Colors: Mud_Green, Cream, Olive
Emotions: satisfaction, accomplishment
Tags: office building, bike rack, arrival, work

Scene 9: The scene shows the man walking into the building,
      greeting colleagues who are also carrying bike helmets
Colors: White, Cream, Olive, Black, Gray
Emotions: friendly, inclusive
Tags: workplace, colleagues, greeting, professional attire

Scene 10: The scene shows the man at his workstation with a
      helmet on his desk, looking out the window at the sunny
      day, hinting at the ride home
Colors: White, Cream, Olive, Gray
Emotions: thoughtful, accomplished
Tags: office, workstation, helmet, window


Select preferred advertisement:
Option 1: A
Option 2: B
Option 3: Both are equally effective


Give reasons for your choice:
_______________________________________
```

9.2 Expert Feedback Collected For Generated Ads

1. Feedback for ad generation for the Maytag Ad shown in Fig 6
    (a) **Expert 1**: "I appreciate the prominent use of the logo in the advertisement. Its placement towards the end, accompanied by a compelling slogan, is in alignment with the brand's advertising strategy."
    (b) **Expert 2**: "In my opinion, the color scheme of the advertisement is stunning. It complements the tone of the advertisement exceptionally well."
    (c) **Expert 3**: "The emotional portrayal in scene 2 could be enhanced. I anticipated a sense of 'recreation' and 'relaxation' to be more effectively conveyed."

2. Feedback for ad generation for the New York Times Ad shown in Fig 5
    (a) **Expert 1**: "One noteworthy aspect in the generated ad description is the concept of 'blocking.' In the ad, the main actor is depicted moving and protesting against various backdrops, including a static background and a subtly shifting frame. This technique is reminiscent of the famous concept utilized in cinematography. While this is not in reflected in the image, I will attribute it to the image generation and not the description generation."
    (b) **Expert 2**: "I like the generated voiceover a lot in terms of story, but I find it hard to fit over the scenes, perhaps this is because the generations dont incorporate transitions/animations."
    (c) **Expert 3**: "I find the overall generated story exceptional in terms of its storytelling in a few ways. 1. The flow of the generated ad, A woman exploring nightlife, protesting, achieving, and nonetheless standing defiant. 2. The slogans are great. 3. The changing head tilt of the woman from sideways to center is a very precise details cinematographer use to paint an overall story or emotion."

3. Feedback for ad generation for the Brainly Ad shown in Fig 4
    (a) **Expert 1**: "I find the overall story formulation to be decent. It portrays kids encountering challenges in solo learning, showcasing easy accessibility and a gradual improvement in confidence and engagement throughout the story. I would still prefer a scene where the UI of the app is somehow shown to the user.[8]"
    (b) **Expert 2**: "I like the use of animated scenes, but I find the incorporation of different main characters slightly jarring. Either they should have been in a common scene, or the main character should not change with every scene. The standout feature of the ad is the utilization of color themes and their harmonization with the emotional tone of each scene."
    (c) **Expert 3**: "Having created Ed-Tech advertisements, I find the storytelling to be excellent. This ad is very persuasive, although it lacks novelty, I still find it to be effective."

10 Annotation Protocol and Participant Details for the LTM Study

Figure 9 shows a visualization of the annotation protocol we followed.

The participants in the study were students who were offered optional course credit and freebies like eatables and a chance to see research and know their memorability scores. The participation was voluntary. The students were shown a protocol of the study and were required to sign the IRB approval, which

---

[8] The generated description of the ad actually shows the student interacting with a visible UI that the image generation model could not respect properly

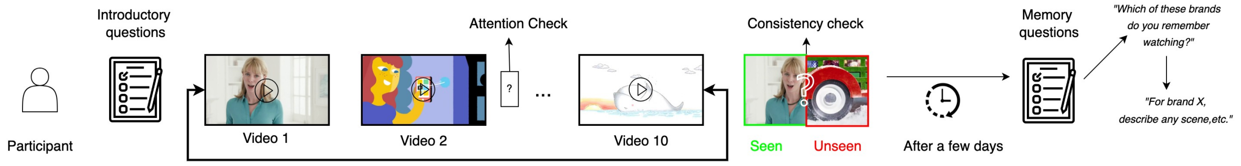


Figure 9: The study protocol we followed for our long term memorability human study. All the previous works follow a game-like annotation protocol, where the study participants compete with each other to get best memorability scores and a participant is excluded from the study if their annotations fall below a certain threshold. We follow a more natural way in which participants fill an initial questionnaire, then watch 10 ads with attention checks on day 1 and in subsequent days, receive a form asking them to fill in what do they remember seeing. Further, using Stable Diffusion, we also ask them to recreate the scenes they remember.

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 were 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 in India. The participants are bilingual and speak a variety of languages, including English. The age range is from 16 to 35 years, and all genders/sexes are encouraged. We saw a roughly 30-70 distribution of females to males.

10.1 Memorability Questionnaire

This section contains the questions we asked before the study, the attention check questions that were asked during the study, and finally, the recognition questions to check which brands were remembered.

10.1.1 Introductory Questionnaire (to be filled before the study starts)

1. I remember seeing ads for the following brands this year:
    - List 15 randomly selected from the list of brands that we have
2. 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)
3. Have you installed any Ad Blocking software in your browser(s)?
    a. Yes
    b. No
4. Do you use a Youtube subscription?
    a. Yes
    b. No
5. 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
6. 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
    - Store Visits
    - Website Ads
    - I primarily search for products

10.1.2 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

10.1.3 Recognition Questions (asked after a few days after watching the videos)

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):

11 Collection of all the Prompts used in the Paper

11.1 GPT-4 Prompts

Listing 11: GPT-4 Prompt to calculate preference between Real Ad (A) and Generated Ad (B)

```
As a seasoned marketer, evaluate the effectiveness of the
      following two ads using a comprehensive set of metrics:

Creativity and Innovation: Originality and uniqueness in
      conveying the message. Use of unexpected ideas or
      elements that grab viewers' attention.

Emotional Connection: Ability to evoke strong, relevant
      emotions in the target audience. Establishing a
      connection between the brand and the viewers' emotions.

Storytelling: Crafting a compelling narrative that engages and
       retains the audience. Creating a memorable experience
      through a coherent and impactful story.

Visual Appeal: Use of strong visual elements, such as striking
       visuals, colors, and graphics. Ensuring that the visual
      elements align with the overall message and brand image.

Brand Alignment: How well the ad aligns with the values,
      mission, and personality of the brand. Consistency with
      the brand's visual identity, tone, and messaging. The ad'
      s ability to leave a lasting impression on viewers
      regarding the brand. Incorporating brand elements that
      make it easy for the audience to remember and recognize.

Target Demographics: Relevance of the ad content and message
      to the target audience. Appropriateness of visuals,
      language, and themes for the specific demographic group.

Based on these criteria, analyze and determine which of the
      two ads is more effective. I will provide you with the
      Voiceover, followed by their scene-by-scene descriptions,
       including the emotions shown in the scene, the text,
      objects, colors, and style of the image.

Ad (A): {Verbalization for Ad (A)}

Ad (B): {Verbalization for Ad (B)}

Give me your answer in a json format, with the following keys:
- ad_a_score: Score between 0 and 10 for Ad A
- ad_b_score Score between 0 and 10 for Ad B
- winner The winner of the two ads
- reason line separated Reasons for the winner in not more
      than 3 lines
```

Listing 12: GPT-4 Prompt to measure consistency of an Ad

```
You are now a seasoned marketer that judges the consistency of
       an advertisement well. The consistency of an Ad can be
      determied by a few metrics (in no particular order) such
      as:
1. Does the voiceover match with the Scenes in the Ad?
2. Do the scene description make a good story?
3. Are the emotions depicted in the scenes consistent with the
       overall ad?
4. Does the ad represent the product and the brand well?

Rate the consistency of the following ad out of 10. Give me
      the rating only and nothing else, or you will be
      penalized.
{Advertisement Description}
```

Listing 13: GPT-4 Prompt to generate ad verbalization with In-Context-Learning (ICL)

```
You are now a seasoned marketer that creates memorable ads
      given its duration, brand and title.
Your output should follow the writing style of the input
      exactly. For example, each scene should look like:
The scene shows {}. The foreground colors of the scene are {},
       and the background colors are {}. The dominant tone of
      the scene is {}. The photography style of the scene is
      {}. The scene has {} visual complexity. The emotions
      shown in the scene are {}.  This scene is categorized by
      the tags {}.
You are only supposed to fill in the {}

Generate the detailed description of a {DURATION_AD1} second
      memorable advertisement titled "{TITLE_AD1}" for the
      brand {BRAND_AD1}
Generate the detailed description of a {DURATION_AD2} second
      memorable advertisement titled "{TITLE_AD2}" for the
      brand {BRAND_AD2}
...
Generate the detailed description of a {DURATION_AD5} second
      memorable advertisement titled "{TITLE_AD5}" for the
      brand {BRAND_AD5}
Generate the detailed description of a {DURATION_TARGET}
      second memorable advertisement titled "{TITLE_TARGET}"
      for the brand {BRAND_TARGET}
```

11.2 Henry Prompts

Given below are the verbalization templates we use to teach Henry and Henry-SEED behavior simulation and content simulation tasks:

Listing 14: Verbalization pattern to predict memorability given advertisement. The same template is used to prompt GPT-3.5, GPT-4, Henry, Henry-Oracle, and Henry-SEED. Note that video tokens are optional.

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

Listing 15: Henry Prompt to generate ad verbalization used to train and evaluate Henry-SEED

```
Generate the detailed description of a {DURATION_TARGET}
    second memorable advertisement titled "{TITLE_TARGET}"
    for the brand {BRAND_TARGET}
```

11.3 Mistral prompt for filtering marketing ads

Listing 16: Mistral Prompt for Ad Filtering

```
"Based on the topic_tags_vocab = {'politics ': 'The art and
    science of governing societies and making decisions that
    affect collective interests.', 'marketing ': 'The process
    of promoting, selling, and distributing products or
    services to consumers, often involving market research,
    advertising, and branding strategies.'} provided, please
    identify the top most relevant topic tag from the
    topic_tags_vocab keys that represent the following
    advertisement based on content and page_name. Please use
    only the most relevant tag and make sure to choose from
    provided topic tags only. Do not include any other tags
    not mentioned in the prompt.Answer with the most relevant
     topic tag only. The advertisement is posted by the page
    Donald J. Trump and has the following content : ['
    President Trump is coming to town! Get your free tickets
    now >>>']. Answer in only politics or marketing."

cleaned_text = "The advertisement is posted by the page {
    page_name} and has the following content : {page_content
    }"
```

12 Computing Infrastructure and Hyperparameters

12.1 Modeling Memorability

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

12.1.1 Images

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

12.1.2 Videos

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

12.2 Generating Memorable Ads

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

12.2.1 Inference hyperparameters

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

13 License and Terms of Release

LAMBDA and UltraLAMBDA are sourced from brand videos from YouTube, Facebook Ads, and CommonCrawl. The dataset annotations and video links contained in 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 do not contain or disclose any identities of their annotators or any specific persons. Since it is handcrafted, LAMBDA makes sure that none of the videos are offensive; UltraLAMBDA being sourced from the internet is noisier. While the videos themselves originate from brands, the 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.

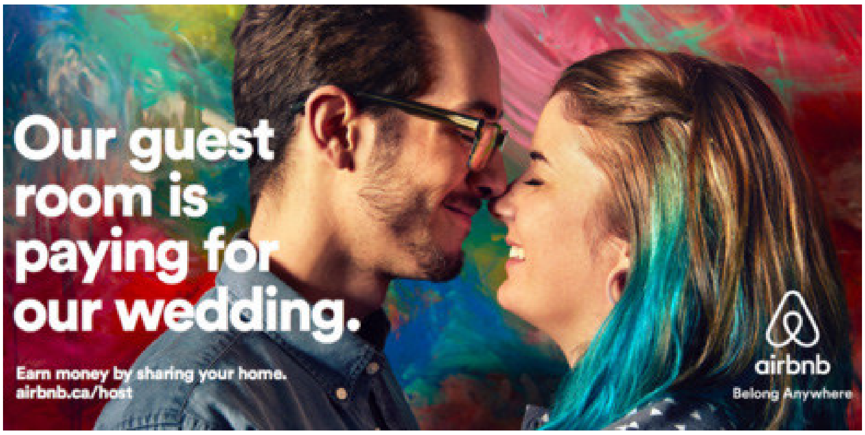

Our guest room is paying for our wedding.
Earn money by sharing your home.
airbnb.ca/host
airbnb
Belong Anywhere

Figure 10: 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.

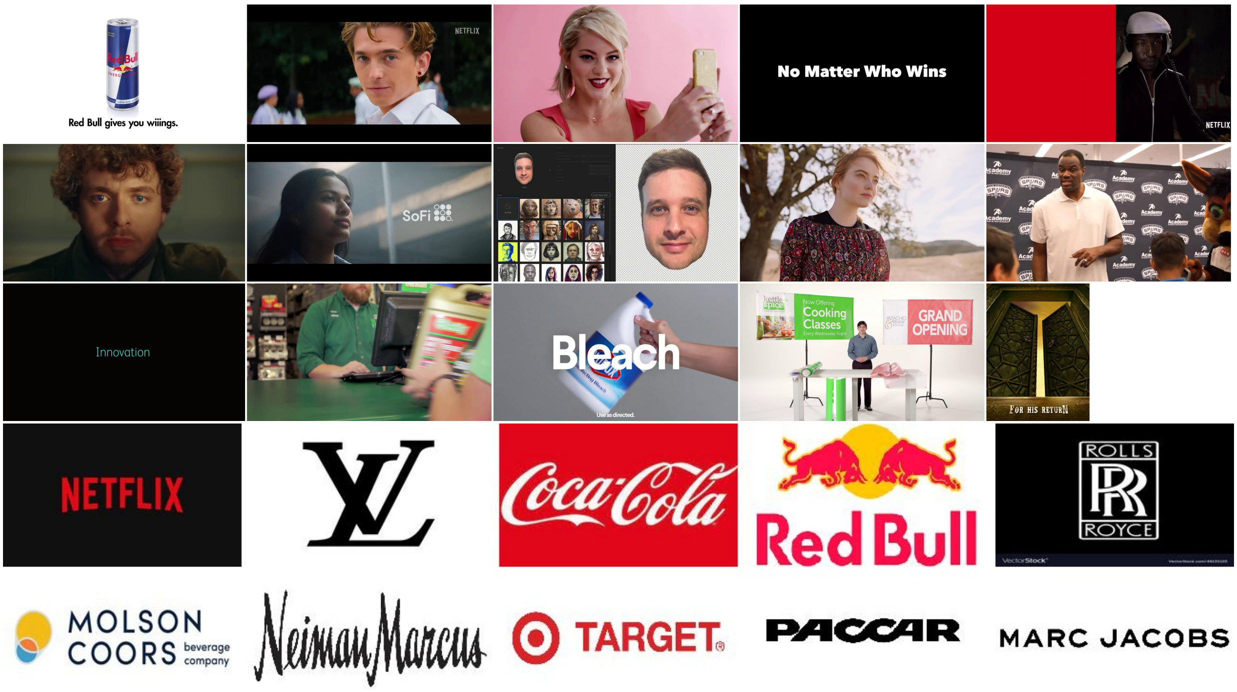


Figure 11: The top three rows show the keyframes from videos in our dataset, LAMBDA, arranged from most to least memorable. The bottom two rows show brands arranged from the most memorable brands to the least.