

# Unsupervised Memorability Modeling from Tip-of-the-Tongue Retrieval Queries

Sree Bhattacharyya<sup>1</sup> Yaman K. Singla<sup>2</sup> Sudhir Yaram<sup>2</sup>  
 Somesh Singh<sup>2</sup> Harini SI<sup>2</sup> James Z. Wang<sup>1</sup>

<sup>1</sup>The Pennsylvania State University <sup>2</sup>Adobe Media and Data Science Research

sfb6038@psu.edu, behavior-in-the-wild@googlegroups.com

## Abstract

*Visual content memorability has intrigued the scientific community for decades, with applications ranging widely, from understanding nuanced aspects of human memory to enhancing content design. A significant challenge in progressing the field lies in the expensive process of collecting memorability annotations from humans. This limits the diversity and scalability of datasets for modeling visual content memorability. Most existing datasets are limited to collecting aggregate memorability scores for visual content, not capturing the nuanced memorability signals present in natural, open-ended recall descriptions. In this work, we introduce the first large-scale unsupervised dataset designed explicitly for modeling visual memorability signals, containing over 82,000 videos, accompanied by descriptive recall data. We leverage tip-of-the-tongue (ToT) retrieval queries from online platforms such as Reddit. We demonstrate that our unsupervised dataset provides rich signals for two memorability-related tasks: recall generation and ToT retrieval. Large vision-language models fine-tuned on our dataset outperform state-of-the-art models such as GPT-4o in generating open-ended memorability descriptions for visual content. We also employ a contrastive training strategy to create the first model capable of performing multimodal ToT retrieval. Our dataset and models present a novel direction, facilitating progress in visual content memorability research.*<sup>1</sup>

## 1. Introduction

In today’s digital age, the constant stream of content delivered through the internet and social media inundates people with more information than ever before. Our minds, however, retain only part of such information effectively. Several factors influence what persists in memory, in-

cluding characteristics of the information itself, as well as individual differences. Studying content memorability, especially for visual media, has long been a research goal within psychology, cognitive science, and computer science. Creating systems that can automatically model content memorability has diverse applications, including in education [12], advertising, marketing [16], and design [39, 48]. Within the field of Artificial Intelligence (AI), applications include creating personalized agents [41], tailored recommender [4], and retrieval systems [44, 46], or providing a better computational understanding of human intelligence [31]. Past research within AI has mainly focused on understanding and modeling visual content memorability, spanning the modalities of images and videos. Adding to the findings from human studies in psychology [8, 33], these computational efforts have found that a large part of memorability can be attributed to factors intrinsic to the visual content [21, 26, 31], such as semantic concepts, emotions, or object categories. This is crucial to understanding and incorporating memorable properties within content, which can potentially generalize for all individuals interacting with it.

While several past research works study visual memorability, they usually depend on obtaining human annotations to create a memorability dataset. The most common method for collecting such data, especially for short-term recall, is in the form of a visual memory game [26]. For experiments studying long-term recall, the process becomes further challenging, as individual participants are required to first view the visual content, and then respond to memorability-related questions a significant amount of time later. In [21], for example, the authors state that the total time for curating the dataset (which contains only 2205 samples) was two years. This makes the data curation process less scalable, both in terms of the number of data points collected and the detail of memorability signals obtained from the participants. For example, most works express memorability for visual content through *singular scores*, aggregated over the data

<sup>1</sup>Code available at: [https://github.com/sreebhattacharyya/web\\_scale\\_memorability](https://github.com/sreebhattacharyya/web_scale_memorability).

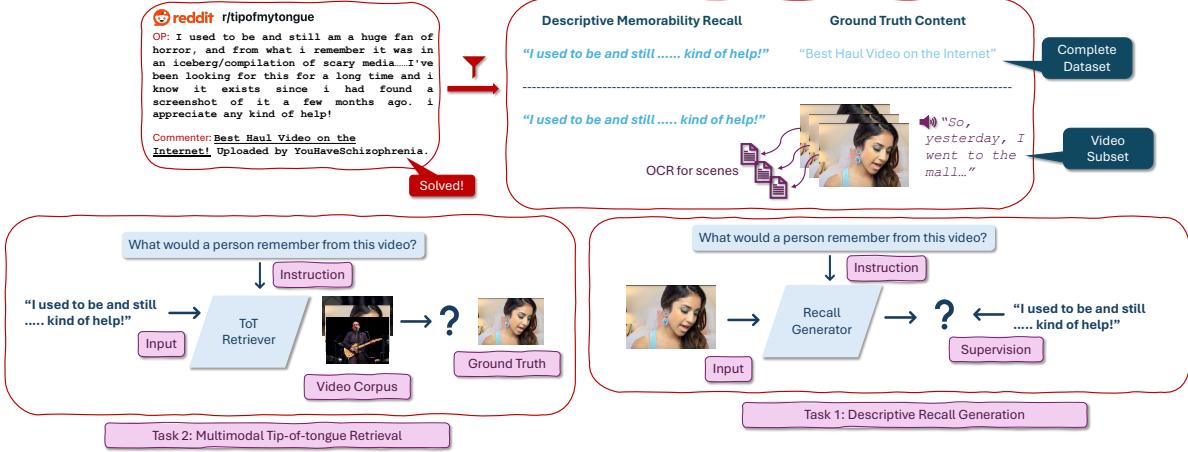


Figure 1. Our complete data collection and task pipeline. We use Tip-of-the-Tongue (ToT) search posts from Reddit (top left), and collect data through a rigorous filtering process. This leads us to obtain data points that are essentially recall-content pairs. The original Reddit search query becomes the descriptive recall, as it is what a user tries to identify the content from their memory recalls, and subsequently expresses on the platform. The correct content is retrieved from within the comments made to the posts. We also create a video-based subset of the data by downloading the raw visual information from YouTube and providing additional details such as audio transcripts and OCR. We propose two tasks using our dataset: Descriptive Recall Generation and Multimodal ToT retrieval. We also present ToT2MEM-RECALL and TOT2MEM-RETRIEVAL, respectively, to generate descriptive memorability recall or perform multimodal retrieval, trained on our dataset.

collected from all participants. This does not account for the rich signals that can be leveraged from **open-ended recall descriptions** of visual content, which can provide directions on exactly **what** is memorable, beyond knowing only **how** memorable something is (recognition).

In our work, we address this gap by harnessing web-scale data from online Tip-of-the-Tongue (ToT) [46] search forums such as Reddit <sup>2</sup>. These online ToT search forums operate on the underlying principle of “wisdom of crowds” [18, 47]: users who are trying to remember some content they have interacted with in the past, but are unable to produce precise identifiers to recall or retrieve them, post inexact descriptions on the forums, while other participants respond with their best guesses for the potential answer. The “correct answer” is finally chosen by the question poser. The advantage of data from such platforms is the availability of descriptive memory recall signals - in the form of the search query posted - as well as the opportunity to collect diverse, large-scale data of free-form interactions. This helps us scale to multiple content domains, such as videos, books, movies, and songs, and long-term descriptive recall for modeling memorability. Additionally, using the “correct answers” present in the comments of such posts, we construct a large-scale video-based subset of our data, which we use to model recall generation and perform ToT retrieval.

Our main contributions can be summarized as:

<sup>2</sup>For example, the Tip Of My Tongue thread on Reddit: <https://www.reddit.com/r/tipofmytongue/>

- We introduce **TOT2MEM**, the first large-scale dataset for modeling visual content memorability with **open-ended recall signals** (Section 3). Our dataset consists of over 470,000 content-recall pairs, spanning multiple content domains and genres. Further, we present a video-based subset of our data, **TOT2MEM-VIDEO**, containing 82,500 video-recall pairs, and the raw visual content (eg., scenes), and audio.
- We provide a detailed analysis of our large-scale dataset, presenting insights on how pervasive content features, like genre, impact memorability (Section 3).
- Using TOT2MEM-VIDEO, we propose the novel task of descriptive recall generation. We present a trained model TOT2MEM-RECALL, capable of generating human-like recall signals for video content. We show that this model generalizes to high-quality out-of-distribution human data for descriptive recall generation and a memorability-related ranking task (Section 4).
- We also create TOT2MEM-RETRIEVAL, using a contrastive training strategy [27] with our dataset, and present the first solution for multimodal ToT retrieval (Section 5).

We visualize our entire data and task framework in Fig. 1. We plan to release all of our data and models publicly.

## 2. Related Work

### 2.1. Visual Content Memorability

A majority of early research efforts in visual content memorability focus on images [7, 7, 19, 26, 29–31], with more re-

Dataset	# Samples	Memorability Signal	Modality	Domain / Type	Data Collection Setting
LaMem	60,000	Recognition	Images	Open / General	Competition-based Experiment
SUN	2,222	Recognition	Images	Open / General	Competition-based Experiment
MemCat	10,000	Recognition	Images	Open / General	Competition-based Experiment
Memento10k	10,000	Recognition	Videos	Single-action amateur videos	Competition-based Experiment
MediaEval	1500	Recognition	Videos	Short clips from Flickr and Twitter	Competition-based Experiment
VideoMem	10,000	Recognition	Videos	Single-action staged video	Competition-based Experiment
LAMBDA	2,205	Recognition and Recall	Videos	Advertisement	Natural Experiment
ToT2MEM	<b>470,000 (total), 82,500 (videos)</b>	<b>Recall</b>	<b>Text, Images, and Videos</b>	<b>General Entertainment</b>	<b>Unsupervised in-the-wild interactions</b>

Table 1. Comparison between current and popular memorability datasets and our dataset.

cent works venturing into studying memorability for video content [11, 21, 32, 38]. Most of these studies model “recognizability” using an aggregate memorability score. Only two past research efforts [7, 21] collect recall-related information. [21] collects descriptive recall signals from participants, between 24 - 72 hours after they have interacted with the videos. However, the specific recall content is not used for modeling memorability. [7], on the other hand, collects recall descriptions from participants, in the presence of a blurred version of the original images in the study. Both of these studies also focus on specific domains - advertisements and visualizations, respectively. Thus, descriptive recall signals have remained relatively underexplored in past works. Further, most of the past studies have also focused on small-scale human data collection, often in a competitive, in-the-lab setting. We compare the scale of our dataset with other popularly used memorability-related datasets in Table 1.

## 2.2. Tip-of-the-Tongue Retrieval

Tip-of-the-Tongue (ToT) Retrieval is a subset of general known-item search [34], where the user searches for an item they have interacted with in the past, while being in a “tip-of-the-tongue” state [46]. Several past works study ToT retrieval by building relevant datasets and benchmarks [1, 5, 28]. Almost all of the existing methods have focused on retrieval in a text-to-text setting, often reducing multimodal content to a textual format (eg, plot descriptions as a proxy for movies) [1]. Most existing datasets are also domain-specific, and relatively limited in scale, while some expand to several casual search domains (eg., books, movie, games, videos, etc.) for qualitative [37] and quantitative analysis [17] of the query content and quality. Several works also present baseline methods for ToT retrieval, including standard methods like BM25 to LLM-based re-ranking of outputs [6]. However, throughout existing literature, the challenging nature of the task has been repeatedly highlighted. [5] show that neither individual parts of a ToT query nor LLM-based reformulations yield good retrieval performance, while [37] find that pre-retrieval indicators cannot predict how long it takes for a query to be solved. Building on these challenges, we introduce the first large-scale dataset and method for multimodal ToT

retrieval, demonstrating that LLM-based embeddings can substantially improve performance.

## 3. ToT2MEM: Using Tip-of-Tongue Retrieval Queries for Descriptive Memorability

In this section, we describe our procedure to collect data and present a detailed analysis of the same.

**Data Filtering and Collection from Reddit:** To collect large-scale memorability data, we focus on threads where Reddit users post descriptions about diverse content items (eg., movie, videos, images, books, quotes, etc.), usually from their memory of having interacted with the item in the past, in a “tip-of-the-tongue” description fashion [46], and pose a question, wanting to know what the exact content item was. Other Reddit users then respond in the comments, usually providing their best guesses of what they think the item could be. We collect data from 2431 unique Reddit threads known for hosting ToT-style searches between January 2017 and June 2022.

We exclude posts that contain NSFW markers or are created by bots. The total number of posts we find after such initial filtering are 1,979,167. We then remove posts that are marked as deleted (about 2% of the total). Out of the remaining posts, 471751 posts are marked as “solved”, meaning that some user was able to post the correct reference to the exact content item being searched. We validate the “solved” status of posts through multiple checks. Firstly, we consider whether an appropriate (Solved) tag appears in the flair CSS class for the post, as is common in many Reddit threads. Then, we iterate through each comment made on the post and identify whether the post author replies explicitly to any of the comments, confirming that the comment provides the correct answer. Further, we also check whether the moderator bot for a given subreddit thread confirms the same correct answer within the post comments. Finally, we verify the “correct answer” obtained through these mechanisms by passing the entire post, along with all comments, to a small LLM (DeepSeek-R1 Distilled on LLaMA 8B) [20] in a few-shot setting, where it independently identifies the correct response by scanning the entire text content of a post. In Appendix A, we provide further details of the

validation process and additional data statistics.

We focus specifically on solved posts, as they show that the recall descriptions provided in the original post were *reliable enough* to find the correct answer. Our dataset comprises all 470,000 solved posts, spanning multiple domains of content items.

### Creation of the Video-Based Data Subset TOT2MEM-VIDEO:

We find that YouTube videos are one of the most frequently searched content types on Reddit platforms (Fig. 4b in Appendix C). Leveraging this, we also create a large-scale video dataset, presenting the raw visual information of videos, to be directly used for training models for downstream tasks. We begin by extracting all YouTube links embedded within posts. We discard posts where the search query does not contain a substantial text description (e.g., consists of only a video link in the query itself, or any media in any other modality, such as audio or image). Then, we filter out posts for which the correct answer is a video hosted in a domain other than YouTube.

We further filter posts to remove private or deleted videos, and only retain videos that are less than 600 seconds in length. As an additional verification mechanism, we utilize DeepSeek-R1 LLaMa 8B [20] in an LLM-as-a-judge setting to obtain the name of the correct video from each post, while also cross-validating with metadata obtained for each video from YouTube. We also manually verify about 100 such responses to ensure that they match and do not find any errors. Following these stages of filtering, we are left with 82,500 recall-video pairs that form TOT2MEM-VIDEO.

For each video in the dataset, we then sample scenes using a context-aware adaptive scene detector<sup>3</sup>. We also detect texts displayed in the sampled scenes through automatic OCR [13] and generate audio transcripts for videos using Whisper [42]. Then, using the OCR texts for each scene, we create a de-duplicated set of scenes for each video, including only consecutive scenes where the OCR changes. This helps reduce both the computational complexity of processing the visual information and contextual noise in the visual data. For additional details on the filtering process, see Appendix B.

### Connecting TOT2MEM with External Content Factors:

Several past works have shown the influence of intrinsic content factors on memorability, such as emotions displayed, popularity [21], or the presence of certain objects [38]. We complement this with an analysis of more pervasive content-related factors, such as genre popularity. We present our findings as follows:

**(a) Popular Items Are Searched More Often and Retrieved Faster:** We use views on YouTube (for videos,

movies, trailers, songs, etc.) as a proxy for external popularity, and find (Fig. 2a), that external popularity is positively correlated with popularity on Reddit search forums. Although externally popular content items (such as a YouTube video watched many times) can be expected to remain stored in memory (because it is being watched so frequently), they may appear in ToT searches because users from different cultural or linguistic backgrounds might lack the exact vocabulary to find it directly. On the contrary, we find a weak negative correlation between external popularity and the time required for users on Reddit to come up with the correct answers to searches (Fig. 2b). Further, we also compare genre-wise popularity (using the average box office collections of each genre in North America across 2017-2022), with the number of searches made about posts belonging to each genre, in Fig. 2c. We find that although Action and Adventure are popular in terms of revenue, Drama is the genre searched for the most.

**(b) Genre-Specific Analysis:** We then take a deeper look at genre-specific properties of the posts. Firstly, we study how quickly, on average, posts from each genre get solved. As demonstrated in Fig. 3a, searches for content belonging to all popular genres get solved quickly, while searches for more niche content, such as documentaries, take the longest time. Second, we study how the number of searches about content within a specific genre change, with respect to time elapsed since its initial release. Most posts on ToT forums are created a long time after the original content is released, with most genres seeing a spike around 8 - 10 years after release. Content belonging especially to popular genres like Adventure, Drama, or Horror is particularly remembered (or attempted to be remembered), even long after it was created.

To explore whether our unsupervised dataset (in particular TOT2MEM-VIDEO) can suffice for modeling content memorability, we propose two novel tasks:

- **Descriptive Recall Generation**, where we train a model using our dataset, to generate human-like memorability signals in the form of open-ended text.
- **Tip-of-the-Tongue (ToT) Retrieval**, where we explore whether the functionality served by online ToT forums (like Reddit) can be replicated automatically, through a model trained using our data.

## 4. Task 1: Descriptive Recall Generation

The first task we propose to benchmark our dataset on, deals with predicting or generating, in a human-like manner, descriptive recall statements for visual content. In this section, we describe the proposed task setting, the experimental setup, and results for recall generation.

**Task Setup.** As described in Section 3, our video-based data subset contains the following: the raw visual content

<sup>3</sup>AdaptiveDetector - SceneDetect

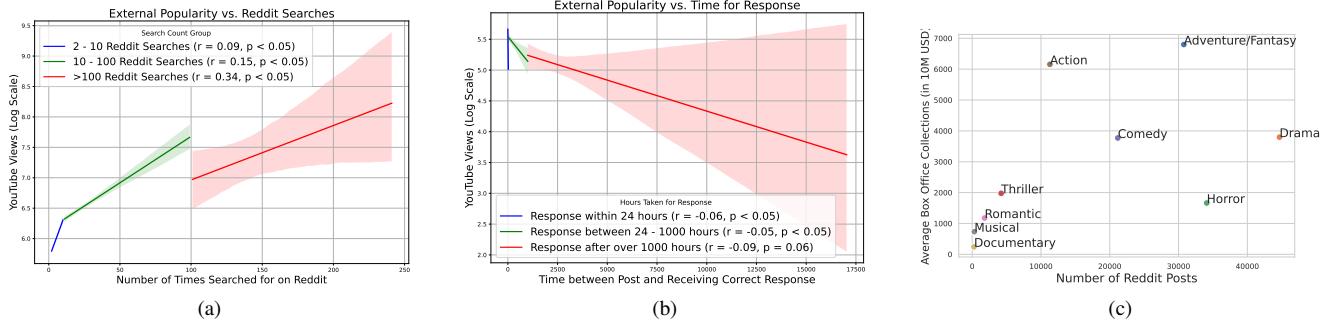


Figure 2. (a) Correlation between external popularity (measured using YouTube views) and number of times searched on Reddit. Zoomed-in graphs for each group (based on the search count) are included in the Appendix C. We also include the analysis using Wikipedia page views as a measure of popularity in the Appendix C. (b) Correlation between external popularity (YouTube views) of the content and the time between the original post about the content is made, and when someone comments with the correct answer. (c) Relationship between popularity, as measured by the average box office collections, of different genres, and the number of posts found in ToT forums. A more detailed picture of the relationship between genre popularity and searches is presented in Appendix C.

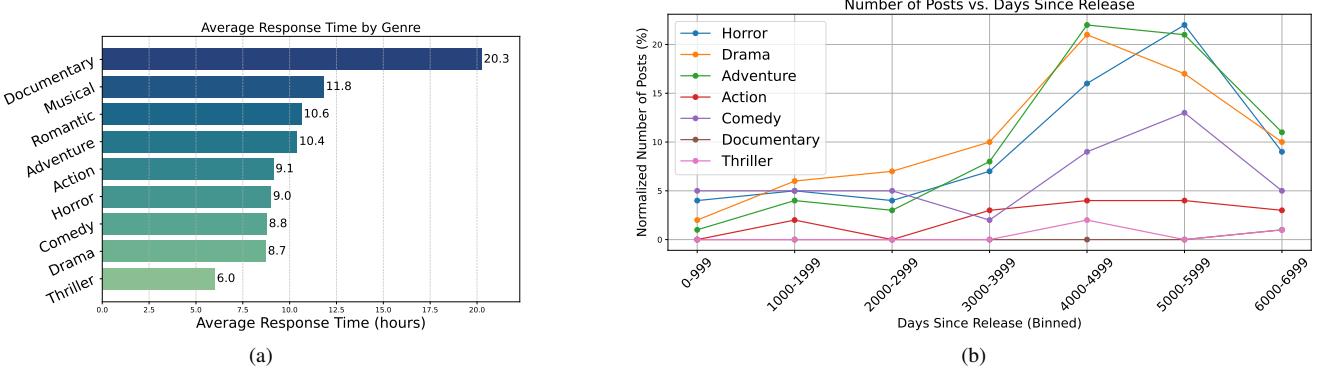


Figure 3. (a) Average Response Time (in Hours) for content belonging to each genre. Response time refers to the time elapsed between a post/search being made and the time when the correct answer is provided in the comments. (b) Comparison of searches made for content in each genre, with time (in days) since the release of that content (as obtained from the Wikipedia page creation date).

(scene images), the corresponding OCR for each scene, and an audio transcript for the entire video. With this, there also exists the ground truth descriptive recall text, obtained from the original Reddit query<sup>4</sup>. The goal of the recall generation task then is: given any visual media  $V$  and the associated information in the form of input, and a task instruction  $I$  (constant across all input samples), a given model  $M(\theta)$  should predict a sequence  $S'$  that describes what a person might remember from the given visual content, using the original ground truth recall  $S$  as the source of supervision. Thus, we simply have,  $M(V, I; \theta) = S$ . We consider a pre-trained Vision-language Model (VLM) as our initial model  $M(\theta)$ .

**Training and Evaluation.** We fine-tune the QwenVL 2.5

<sup>4</sup>For training, we include all sentences of the original recall signals except ones that describe purely episodic memory to mitigate hallucinated responses about episodic memory.

7B model [2], in a parameter-efficient manner, using LoRA [24], on ToT2MEM-VIDEO. To show the effectiveness of the use of our dataset, we compare our results with various state-of-the-art multimodal baselines, capable of dealing with multi-image or video inputs, evaluating them in a zero-shot manner. Precisely, among open-source models, we evaluate InternVL 2.5 8B [10], QwenVL 2.5 7B [2], and Qwen Omni 2.5 7B [52] with multi-image inputs in a zero-shot manner, while also comparing with InternVideo-2.5 8B [49] to which we provide direct video inputs. Among proprietary models, we evaluate GPT4-o [25]. The baselines are chosen primarily owing to their superior performance in general multimodal and video understanding tasks [22, 35]. Zero-shot evaluation is chosen for the baselines, as the inputs already contain multiple images; providing few-shot examples becomes further confusing (especially for open-source models) and expensive. We utilize popular text generation metrics from

Model	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
GPT-4o (zero-shot)	0.193	<u>0.197</u>	0.204	0.027	0.117	0.81
InternVL-2.5 8B (zero-shot)	<u>0.211</u>	0.196	<u>0.216</u>	0.059	<u>0.148</u>	<u>0.82</u>
InternVideo-2.5 8B (zero-shot)	0.189	0.159	0.196	0.033	0.128	<u>0.82</u>
Qwen-2 VL 7B (zero-shot)	0.173	0.192	0.024	<u>0.123</u>	0.122	<u>0.82</u>
Qwen-2.5 VL 7B (zero-shot)	0.196	0.153	0.2	0.021	0.126	<u>0.82</u>
Qwen-2.5 Omni 7B (zero-shot)	0.083	0.173	0.12	0.024	0.072	<u>0.82</u>
<b>TOT2MEM-RECALL</b>	<b>0.242</b>	<b>0.293</b>	<b>0.304</b>	<b>0.152</b>	<b>0.251</b>	<b>0.85</b>

Table 2. Primary results for evaluation on the descriptive recall generation task. The top scores are shown in bold, while the second-highest scores are underlined.

Model	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
GPT-4o (zero-shot)	0.16	0.20	0.177	0.029	0.105	0.823
InternVL 2.5 8B (zero-shot)	0.14	<u>0.19</u>	0.16	0.03	0.11	0.82
Qwen 2 VL 7B (zero-shot)	<u>0.21</u>	0.17	<u>0.22</u>	<u>0.04</u>	<u>0.13</u>	<u>0.84</u>
Qwen 2.5 VL 7B (zero-shot)	0.19	0.18	0.2	0.03	0.12	0.83
Qwen 2.5 Omni 7B (zero-shot)	0.04	0.12	0.07	0.02	0.05	0.79
<b>TOT2MEM-RECALL</b> (zero-shot)	<b>0.27</b>	<b>0.37</b>	<b>0.29</b>	<b>0.14</b>	<b>0.24</b>	<b>0.86</b>

Table 3. Results on the LAMBDA Recall Dataset [21]. The topmost scores are shown in bold, and the second-highest scores are underlined.

the literature, such as BLEU [40] and ROUGE [36], and also include METEOR [3] and BERTScore [14] to capture semantic similarity.

**Results.** The primary results are presented in Table 2. Our model *outperforms the compared baselines on all of the metrics*, demonstrating that our dataset helps capture the nuances of open-ended memorability descriptions. Importantly, our model is based on an unmodified Qwen2.5-VL-7B backbone and differs from the baselines only in being fine-tuned on our dataset. This setup ensures that the comparison directly reflects the value of our dataset rather than architectural differences <sup>5</sup>. It also demonstrates how general multimodal reasoning ability and world knowledge, which models like GPT4-o excel at, might not be sufficient for predicting recall signals, given its unintuitive nature [31]. We also assess the statistical significance of the improvement of TOT2MEM-RECALL on the top-performing baseline (InternVL 2.5 8B) using Wilcoxon’s test [50], and find that the improvements are highly statistically significant ( $p << 0.05$ ). For additional experiments, including ablations, see Appendix E. We also provide qualitative examples and failure analysis in Appendix F.

**Zero-shot Transfer.** To understand whether the performance gains simply stem from having a distributional advantage, we further evaluate the baselines and our model

<sup>5</sup>We show, similarly, the effectiveness of using our dataset to tune another strong baseline, InternVL 8B, in Appendix E.

Model	Accuracy (%)
InternVL-2.5 8B	18.3
Qwen 2 VL 7B	79.28
<b>TOT2MEM-RECALL</b>	82.04

Table 4. The ranking@1 accuracy of our model and two other strong baselines on the Prompt Recall Ranking Task. All models are evaluated zero-shot.

on an unseen dataset: LAMBDA [21]. LAMBDA is the only other known descriptive recall dataset, containing a total of 2205 samples. The high-quality recall signals in this dataset are collected from human participants through a supervised lab experiment. We show that our model, despite being trained on our unsupervised dataset, generalizes zero-shot and again outperforms all other baselines (Table 3). This implies that the memorability signals captured from TOT2MEM-VIDEO are general, relevant, and aligned with high-quality recall signals provided by humans.

**Human Evaluation.** To further validate the robustness of our dataset, we examine whether TOT2MEM-RECALL can generalize to tasks beyond its training setup using high-quality human data collected in a supervised setting. Building on the recognition task studied in [21], we introduce a *Prompt Recall Ranking (PRR)* task. We conduct a human study with 163 participants from an academic institution. Participants were shown the original videos from

the LAMBDA dataset, and after a 48-hour delay, were instructed to recreate the most memorable scene from a single advertisement video using a generative model (Stable Diffusion v2.1) [45]. Each participant provided natural-language prompts corresponding to their memory of the advertisement, iteratively refining up to five prompt variations per scene, resulting in a total of 815 prompts. Additional details about the data collection process are presented in Appendix G. This procedure yielded a dataset linking (a) the original advertisement videos from LAMBDA [21]<sup>6</sup>, (b) human-generated prompts for reconstructing memorable scenes (referred to as diffusion prompts), and (c) the final images generated from these prompts using Stable Diffusion.

In the PRR task, the model is given an original video along with five human-generated diffusion prompts—only one of which is correct—and must identify the correct corresponding prompt. Results for this evaluation are shown in Table 4. We find that TOT2MEM-RECALL, although trained solely on our unsupervised dataset and for the distinct task of descriptive recall generation, generalizes well to the PRR task and significantly outperforms strong zero-shot baselines such as InternVL and Qwen 2 VL<sup>7</sup>. This demonstrates that training on TOT2MEM helps models acquire generalizable memorability signals that transfer effectively to human-curated data, even in a distinct, previously unseen task setting.

## 5. Task 2: Tip-of-the-Tongue Retrieval

Tip-of-the-tongue (ToT) retrieval differs from standard retrieval tasks in that the search queries are often inexact, vague, or even misleading. Using our dataset, we introduce the novel task of *multimodal ToT retrieval* and present a model designed for this purpose, TOT2MEM-RETRIEVAL.

**Task Setup.** Similar to traditional retrieval settings, the task involves searching within a set of documents  $D = \{d_1, d_2, \dots, d_n\}$  to find the most relevant document  $d_i$  for a given query  $q$ . In our case,  $q$  is the noisy textual description provided by a user, the document corpus is a set of videos, and the target document is the correct video the user is attempting to recall. To enable retrieval, both queries and documents must be mapped into a shared embedding space. We achieve this by adapting a vision–language model (VLM) and using its last-layer hidden state representations.

Formally, each training instance consists of a query–video pair  $(q, t^+)$ , where  $t^+$  is the ground-truth video, and a set of hard negative videos  $t^-$  mined based on semantic similarity (details in the Supplementary). Following [27], we prepend a task-specific instruction to the queries. Each training query input ( $q$ ) thus combines:

<sup>6</sup>Note that the original LAMBDA dataset is also curated in a supervised lab setting.

<sup>7</sup>For an additional discussion on the performance gap between the baselines, see Appendix G

(a) the task instruction, (b) the video descriptions (transcripts and OCR text), and (c) sampled scene images from the video. The training target is the associated recall description ( $t$ ). Note that this training configuration reverses the evaluation setting, where recall text is used as the query and videos are the documents.

We obtain last-layer representations  $(h_q, h_{t^+})$  for  $(q, t^+)$  and  $(h_{t^-})$  for negatives  $t^-$ , and optimize the standard InfoNCE loss [27]:

$$\min \mathcal{L} = -\log \left( \frac{\phi(h_q^{\text{inst}}, h_{t^+})}{\phi(h_q^{\text{inst}}, h_{t^+}) + \sum_{t^- \in \mathcal{N}} \phi(h_q^{\text{inst}}, h_{t^-})} \right),$$

where  $\mathcal{N}$  includes both hard negatives and in-batch negatives. After training, the adapted embedding model is used to represent both the queries and the document corpus in the same space for retrieval.

**Evaluation Setup.** We evaluate several baselines for the retrieval task to compare with our model. Across all setups, the query is fixed as the ground-truth ToT search description provided by the user, while the baselines vary in how they represent the document corpus. Below, we describe each setting and the models used:

- **Ground Truth Recall as the Document Set (Topline):**

We explore what the best possible performance is for the ToT retrieval task. For this, we use the *ground truth recall statements* themselves as the document corpus to serve as a proxy for the respective videos. In other words, the set of queries  $Q = \{q_1, q_2, \dots, q_n\}$ , and the document corpus,  $D = \{d_1, d_2, \dots, d_n\}$  are *identical*. Here, the retrieval task reduces to a text-to-text retrieval setting. As both of the queries and documents are in the text format here, we use BGE Embeddings [9] to embed the queries and documents into the same space.

- **Generated Recall as the Document Set:** Similar to the topline, in this setting, we use *recall statements generated by models* as the document corpus, where each generated recall statement would serve as a proxy for the respective videos. We use descriptive recall signals for each video in our dataset, generated in a zero-shot manner by InternVL 2.5 8B [10], as well as TOT2MEM-RECALL from Section 4. In this setting, too, the task is reduced to text-to-text retrieval. Similar to the topline setting, we also embed the query (ground truth recall statements) and documents (generated recall statements), both of which are text-based, using BGE [9] to then use the embeddings for similarity-based search and retrieval.

- **Original Videos as the Document Set:** This setting represents the original text-to-video retrieval task. We compare with several baseline embedding models, such as GME [54], and VLM2Vec-V2 [27]. For our contrastive training strategy, we use VLM2Vec-V2 as the initial model, fine-tuning the unmodified architecture with TOT2MEM-VIDEO. All of these embedding models

Embedding Model	Document Set	Effective Task	Recall/MRR@1	@10	@100
BGE	Query Set	T2T Retrieval	96.5/96.5	98.7/97.4	99.6/97.4
BGE	Generated Recall by InternVL	T2T Retrieval	6.6/6.6	17.1/9.4	37.1/10.1
BGE	Generated Recall by ToT2MEM-RECALL	T2T Retrieval	9.1/9.1	21.2/12.6	43/13.3
GME-QwenVL	Original Videos	T2V Retrieval	5.7/5.7	17.6/9	41.6/9.8
GME-QwenVL w/ Instruction	Original Videos	T2V Retrieval	7.1/ 7.1	20.4/10.9	45.5/11.8
VLM2Vec-V2.0	Original Videos	T2V Retrieval	<u>10.02/10.02</u>	<u>25.4/14.5</u>	<u>48.5/15.3</u>
TOT2MEM-RETRIEVAL	Original Videos	T2V Retrieval	<b>10.9/10.9</b>	<b>27.5/15.7</b>	<b>50.6/16.5</b>

Table 5. Tip-of-the-tongue retrieval evaluation. The topmost performance is shown in bold, and the second-highest scores are underlined. T2T retrieval stands for Text-to-Text Retrieval, and T2V retrieval stands for Text-to-Video Retrieval.

are used simply to embed both queries (text recall statements) and documents (all videos) into the same embedding space, followed by a similarity matching-based retrieval process.

For all of these settings, we report results on standard metrics used in retrieval - Recall@k and MRR@k.

**Results.** Our results in Table 5 highlight both the difficulty of the task and the value of contrastive training on our dataset. First, the topline setting achieves only near-perfect scores, even though queries and documents are identical, showing how semantically similar user ToT queries can be.

Second, we find that recall statements generated by our model (Section 4) serve as strong proxies for the corresponding videos. Retrieval performance using our generated recalls surpasses that of other state-of-the-art generators, and remains comparable to embedding-based methods such as GME that leverage the full video. This indicates that our recall signals effectively capture memorability-related information and can substitute for visual content.

Finally, our contrastively trained model <sup>8</sup> outperforms all baselines across metrics. While its scores fall short of the topline, they are competitive in the broader ToT retrieval literature. Notably, even state-of-the-art methods for purely textual ToT retrieval – a much simpler task – achieve Recall@10 only in the 3–30% range [1, 5, 6]. These results position our approach as an early but meaningful step toward robust multimodal ToT applications.

## 6. Discussion and Conclusion

**Data Quality, Noise, and Ethical Considerations.** We present the first large-scale dataset for modeling descriptive memorability signals, built from online tip-of-the-tongue (ToT) search platforms. To demonstrate its utility, we introduce two new tasks: descriptive recall generation and multimodal ToT retrieval. Training on this unsupervised web-scale dataset yields substantial gains over state-of-the-art baselines and enables fine-tuned models to generalize to

<sup>8</sup>Built on the strongest zero-shot baseline, VLM2Vec, fine-tuned with our dataset.

two human datasets collected in controlled lab settings, including a novel zero-shot ranking task.

While its scale inevitably introduces minor noise - such as occasional errors in solved-answer detection, residual episodic signals, or OCR/ASR recognition errors - the strong empirical gains achieved on multiple benchmarks indicate that such noise does not hinder performance. This suggests the dataset provides high-quality supervisory signals for modeling long-term memorability. In future work, it could also support controlled studies on how large vision-language models handle noisy or ambiguous inputs, including whether such noise can induce hallucinations.

The dataset may also reflect the demographic and topical biases of Reddit, such as the dominant presence of entertainment-related content, searched for in English. To mitigate ethical risks, we release only post IDs, URLs, and derived features (frames, transcripts, OCR) without user identifiers, adhering to Reddit’s terms of service and fair-use principles.

**Future Directions.** Our work opens up avenues for studying nuanced memorability signals and improved applications in practically relevant downstream tasks like tip-of-the-tongue retrieval. Several potential future directions exist for each of the tasks proposed. Using the detailed recall descriptions, interventions could be created for generating personalized and memorable content, providing an opportunity for applications in advertising. At the same time, exploration of ethical concerns with such applications remains critical to ensure users are not manipulated adversely for commercial gains. An empirical analysis of factors affecting memorability (e.g., emotions, episodic context, etc.) could also provide potential insights into the workings of human long-term memory. For multimodal ToT retrieval, advanced feature representations can be crafted or learned using memorability-specific modules. Scaling to the web level to ensure that multimodal ToT search can actually be performed, agentic search systems can also be used to solve the cold-start problem in retrieval. We hope that the public release of our dataset and methods can help pave the path for large-scale applications in the future.

## References

- [1] Jaime Arguello, Adam Ferguson, Emery Fine, Bhaskar Mitra, Hamed Zamani, and Fernando Diaz. Tip of the tongue known-item retrieval: A case study in movie identification. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 5–14, 2021. 3, 8
- [2] Shuai Bai, Kebin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5, 17
- [3] Satanjeev Banerjee and Alon Lavie. Meteor: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005. 6
- [4] Amin Beheshti, Shahpar Yakhchi, Salman Mousaeirad, Seyed Mohsen Ghafari, Srinivasa Reddy Goluguri, and Mohammad Amin Edrisi. Towards cognitive recommender systems. *Algorithms*, 13(8):176, 2020. 1
- [5] Samarth Bhargav, Anne Schuth, and Claudia Hauff. When the music stops: Tip-of-the-tongue retrieval for music. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2506–2510, 2023. 3, 8
- [6] Luís Borges, Rohan Jha, Jamie Callan, and Bruno Martins. Generalizable tip-of-the-tongue retrieval with llm re-ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2437–2441, New York, NY, USA, 2024. Association for Computing Machinery. 3, 8
- [7] Michelle A. Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S. Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. Beyond memorability: Visualization recognition and recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):519–528, 2016. 2, 3
- [8] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual Long-term Memory has a Massive Storage Capacity for Object Details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008. 1
- [9] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024. 7
- [10] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muhan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 5, 7
- [11] Romain Cohendet, Claire-Hélène Demarty, Ngoc QK Duong, and Martin Engilberge. Videomem: Constructing, analyzing, predicting short-term and long-term video memorability. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2531–2540, 2019. 3
- [12] Nelson Cowan. Working Memory Underpins Cognitive Development, Learning, and Education. *Educational Psychology Review*, 26(2):197–223, 2014. 1
- [13] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. Paddleocr 3.0 technical report, 2025. 4
- [14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 6
- [15] David Elsweiler, Ian Ruthven, and Christopher Jones. Towards memory supporting personal information management tools. *Journal of the American Society for Information Science and Technology*, 58(7):924–946, 2007. 12
- [16] Marian Friestad and Esther Thorson. Emotion-eliciting Advertising: Effects on Long-Term Memory and Judgment. *Advances in Consumer Research*, 13(1), 1986. 1
- [17] Maik Fröbe, Eric Oliver Schmidt, and Matthias Hagen. A large-scale dataset for known-item question performance prediction. In *QPP++@ ECIR (pp. 13-19)*, 2023. 3
- [18] Francis Galton. Vox populi, 1907. 2
- [19] Lore Goetschalckx and Johan Wagemans. Memcat: a new category-based image set quantified on memorability. *PeerJ*, 7:e8169, 2019. 2
- [20] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3, 4, 13
- [21] SI Harini, Somesh Singh, Yaman K Singla, Aanisha Battacharyya, Veeky Baths, Changyou Chen, Rajiv Ratn Shah, and Balaji Krishnamurthy. Long-term ad memorability: Understanding & generating memorable ads. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5707–5718. IEEE, 2025. 1, 3, 4, 6, 7, 18, 20
- [22] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020. 5
- [23] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. 19
- [24] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 5, 13
- [25] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5

- [26] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR 2011*, pages 145–152. IEEE, 2011. 1, 2
- [27] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *ICLR*, 2025. 2, 7
- [28] Ida Kathrine Hammeleff Jørgensen and Toine Bogers. “kinda like the sims... but with ghosts?”: A qualitative analysis of video game re-finding requests on reddit. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, pages 1–4, 2020. 3
- [29] Aditya Khosla, Jianxiong Xiao, Phillip Isola, Antonio Torralba, and Aude Oliva. Image memorability and visual inception. In *SIGGRAPH Asia 2012 Technical Briefs*, pages 1–4. 2012. 2
- [30] Aditya Khosla, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Memorability of image regions. *Advances in neural information processing systems*, 25, 2012.
- [31] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, pages 2390–2398, 2015. 1, 2, 6
- [32] Rukiye Savran Kiziltepe, Lorin Sweeney, Mihai Gabriel Constantin, Faiyaz Doctor, Alba García Seco de Herrera, Claire-Hélène Demarty, Graham Healy, Bogdan Ionescu, and Alan F Smeaton. An annotated video dataset for computing video memorability. *Data in Brief*, 39:107671, 2021. 3
- [33] Talia Konkle, Timothy F Brady, George A Alvarez, and Aude Oliva. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of experimental Psychology: general*, 139(3):558, 2010. 1
- [34] Jin Ha Lee, Allen Renear, and Linda C Smith. Known-item search: Variations on a concept. *Proceedings of the american society for information science and technology*, 43(1):1–17, 2006. 3
- [35] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 5
- [36] CY LIN. Rouge: A package for automatic evaluation of summaries. In *Proc. Workshop on Text Summariation Branches Out, Post-Conference Workshop of ACL 2004*, 2004. 6
- [37] Florian Meier, Toine Bogers, Maria Gäde, and Line Ebdrup Thomsen. Towards understanding complex known-item requests on reddit. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 143–154, 2021. 3
- [38] Anelise Newman, Camilo Fosco, Vincent Casser, Allen Lee, Barry McNamara, and Aude Oliva. Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *European Conference on Computer Vision*, pages 223–240. Springer, 2020. 3, 4
- [39] Dale Owen. Designing with memory in mind. 1
- [40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6
- [41] Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023. 1
- [42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 4
- [43] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019. 16
- [44] Henry L Roediger. Why retrieval is the key process in understanding human memory. In *Memory, Consciousness and the Brain*, pages 52–75. Psychology Press, 2013. 1
- [45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 7, 20
- [46] Bennett L. Schwartz and Janet Metcalfe. Tip-of-the-tongue (tot) states: Retrieval, behavior, and experience. *Memory & Cognition*, 39(5):737–749, 2011. 1, 2, 3
- [47] James Surowiecki. *The wisdom of crowds*. Vintage, 2005. 2
- [48] Julien Venni and Mireille Bétrancourt. Aesthetics in Hypermedia: Impact of Colour Harmony on Implicit Memory and User Experience. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, page 215–219, New York, NY, USA, 2021. Association for Computing Machinery. 1
- [49] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhui Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 5
- [50] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992. 6
- [51] Benjamin Wortman and James Z Wang. Hicem: A high-coverage emotion model for artificial emotional intelligence. *IEEE Transactions on Affective Computing*, 15(3):1136–1152, 2023. 13
- [52] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 5

- [53] Junchi Yao, Shu Yang, Jianhua Xu, Lijie Hu, Mengdi Li, and Di Wang. Understanding the repeat curse in large language models from a feature perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7787–7815, Vienna, Austria, 2025. Association for Computational Linguistics. 19
- [54] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024.

7

## A. Data Collection From Reddit

In this section, we provide additional methodological details and results for our process of collecting data from Reddit.

### A.1. Exploratory Data Statistics

In this section, we provide some additional statistics of the entire TOT2MEM dataset collected from Reddit. Firstly, we show in Fig. 4a the top-10 threads that are dataset is built upon. The top-10 threads contribute to over 90% of the data points in the dataset. Then, in Fig. 4b, we show that some popular search items on these platforms include YouTube videos, games, and movies, which are usually all pertaining to casual or entertainment-related searches. Congruent with this, we also find a large presence of links from the YouTube domain in the comments, as shown in Fig. 4c. We also empirically corroborate the finding that users in tip-of-the-tongue states (thereby, in these online communities) experience frustration more frequently [15], by showing that negative emotions are expressed in these search posts much more often than positive emotions (Fig. 4d).

### A.2. Validation of Correct Answers

As mentioned in the Section 3, we use several parallel pipelines for validating the correct answer obtained from a post. Fig. 5 provides an example of a post from Reddit, where different identifiers of the valid answer can be found. The provided example also shows the case where two videos are linked as the “correct answer”. Using our automated resolution mechanism, we retain both correct answers in case multiple video links are attached to them, as it provides additional nuanced signals for recall. For example, if the same recall query could be related to multiple correct solutions, it could help in learning shared aspects or signals about memorability. Precisely, we use regular expressions to find the presence of links in the solved comments. If multiple links can be resolved, the correct answer is formatted as a list of links. However, if no links are found, and the correct answer consists of only text, the entire text is stored as the solved post.

We also use LLM-based validation, where DeepSeek-R1 Distilled LLaMA 8B is used to resolve the name of the post. We provide the following prompt to the model, which also includes querying for other related information about the content:

**Basic Prompt Template:** "You will be given an online post, where a user asks about some content that they are trying to remember, but can only provide currently a vague description of, from their memory. You will also be provided with the response that is given to the user, which contains the name, or link of the correct content item that the user is trying to find about. Your task is, given this conversation, to respond clearly with the name (or link) and type of the content item that the user is asking about. You are required to answer strictly in a JSON format, as follows: `content name`: "name of the content item and the link to it, if present", `category`: "could be movie, book, youtube video, game, song, quote, or anything", `genre`: "could be anything among comedy, drama, horror, fantasy, adventure, or not applicable", `objects`: "standard object categories", `emotions`: "any common emotions that maybe present". Answer strictly in the JSON format, with the correct content item name, and category for the following user post and reply: ". "Original Reddit post search query" "Correct Answer from the comments"

Through the LLM-based verification process, we find >95% of the posts to have solved answers, coherent to what is found through the rule-based resolution process. Among the mismatched posts, we manually inspect 80 posts and find the response from the rule-based mechanism to be more reliable, as it includes the entire text portion of the solving comment. We thus retain the solution from the automated check as part of the dataset, instead of the resolution found by the LLM.

## B. Video-Based Data Subset

We collect and download all available YouTube videos linked in the comments of posts available in our dataset. Using this, we create a multimodal dataset of visual data-recall signal pairs. We initially find around 120000 YouTube links embedded within the posts or comments. Among these initial YouTube videos, over 85% are relatively shorter in length, under 10 minutes of duration. After filtering out links that are part of the post (the question or the search query), deleted posts, or deleted, gated, or corrupted media, our final dataset consists of 82,500 video-text pairs. The maximum number of de-duplicated scenes allowed for each video in the dataset is 30, to ensure that the entire input can fit into the context window of models, and limit computational complexity. All of the videos contain audio streams, and over 60% of the videos contain at least one scene with a non-null associated OCR text.

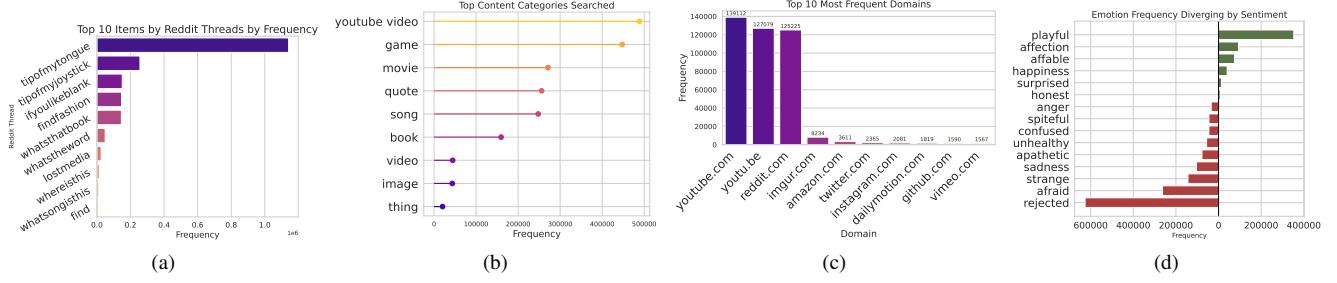


Figure 4. (a): Top 10 Reddit threads used to construct our dataset. (b): Top 10 content types included in our dataset. (c): Top domains to which direct links are present in the dataset, indicating most usually that the correct content item is referred to using a link to these domains. (d): Distribution of emotions in the original Reddit posts. In other words, these are the emotions expressed within the recall signals. The analysis uses the culturally robust HICEM emotion model [51].

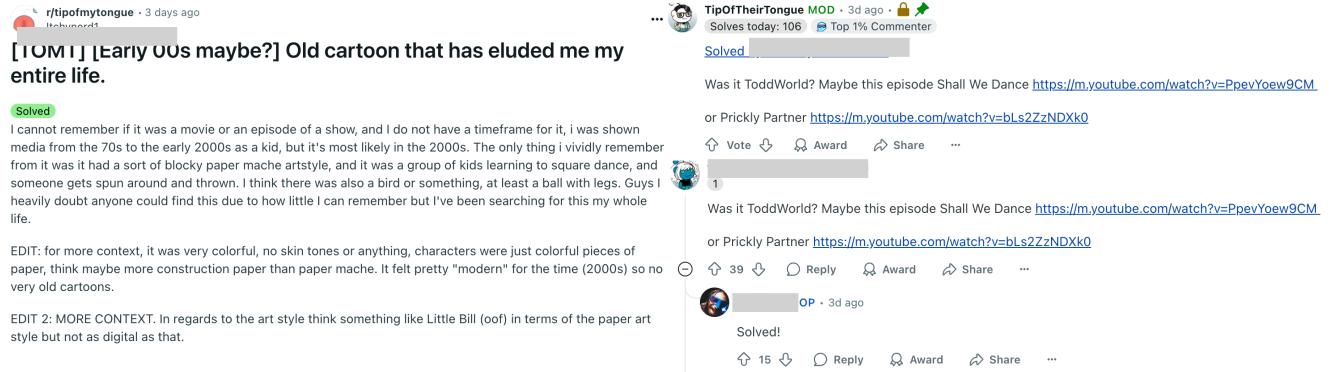


Figure 5. An example of different signals utilized for solving the valid answer. The thread moderator bot provides a comment, which is usually pinned, highlighting the correctly solved answer. Note that the thread moderator usually exactly copies the correct answer and provides it additionally. Further, the actual solving comment can also be found by tracking replies from the original poster, in case it provides directly confirming signals (such as in this case, by saying "Solved!").

## C. Connecting with External Factors

To infer several subjective aspects of each post, such as the content category or genre it belongs to, or different objects, emotions present in it, we use DeepSeek Distilled LLaMA 8B [20], in a few-shot manner. The same prompt, presented above, is used.

Further, we use the APIs of the following websites to collect popularity data: Wikipedia, YouTube, iMDB, and the movie statistics platform The Numbers<sup>9</sup>.

We also study the memory content of the original posts (recall signals) in our dataset. We classify, in an unsupervised manner, each sentence within a Reddit post using DeepSeek-Distilled LLaMA 8B [20]. Sentences within a post may either describe content-related memory, describing, for example, a video an individual may have interacted with, episodic memory, providing additional context about such an interaction, or neither. Further, content-related memory sentences may describe semantic information about the content (eg., plot of a movie) or non-semantic

information (eg., visual elements, location, release time, or actors in a movie). Using a few-shot approach, we tag sentences of each original post, finding that about 57% of all sentences describe content-related memory, while 16% describe personal, episodic memory, and the rest cannot be classified into either category. Within content-related information, the majority of sentences (68%) describe non-semantic information, while the rest provide semantic descriptions of the content.

We include several additional analysis results, complementing the genre-specific and popularity-related analysis, and include them in Figures 6 to 13.

## D. Details about Implementation

**Model Training.** Our video-based dataset is split into training and test sets, containing 80,000 and 2500 samples, respectively. For both the recall generation and retrieval task, our model is trained using Low Rank Adaptation [24], with a rank of 64, while keeping the vision model entirely frozen. The training is completed on 4 H100-80 GPUs, and

<sup>9</sup><https://www.the-numbers.com/>

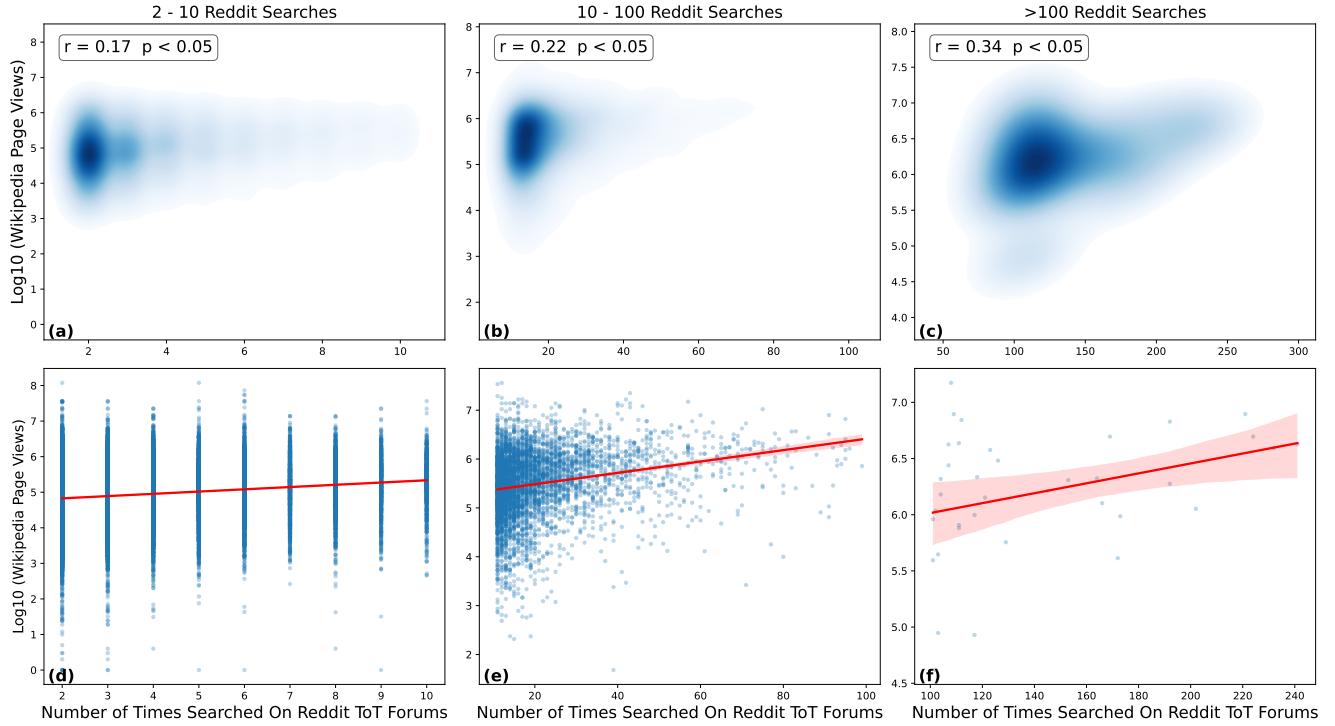


Figure 6. Correlation of Wikipedia-based popularity and number of searches on Reddit, for any given content item, shown in different groups, based on the number of searches made.

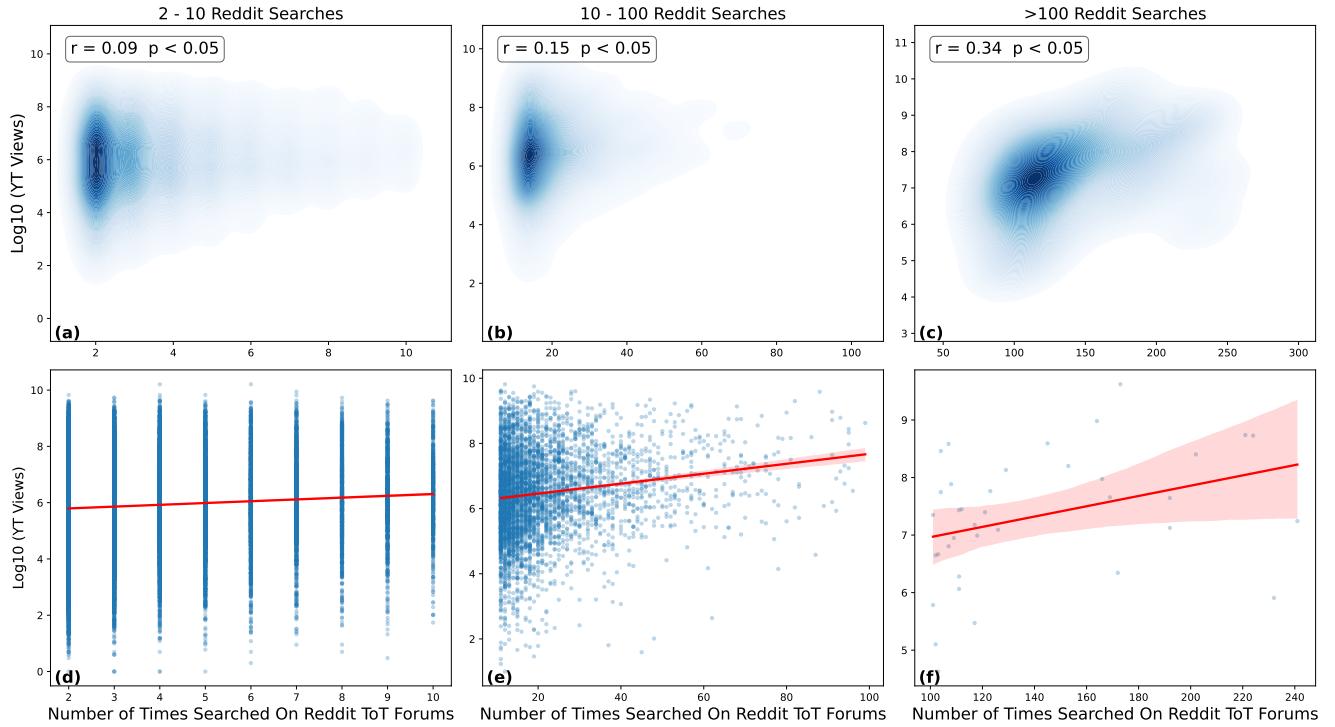


Figure 7. Correlation of YouTube-based popularity and number of searches on Reddit, for any given content item, shown in different groups, based on the number of searches made.

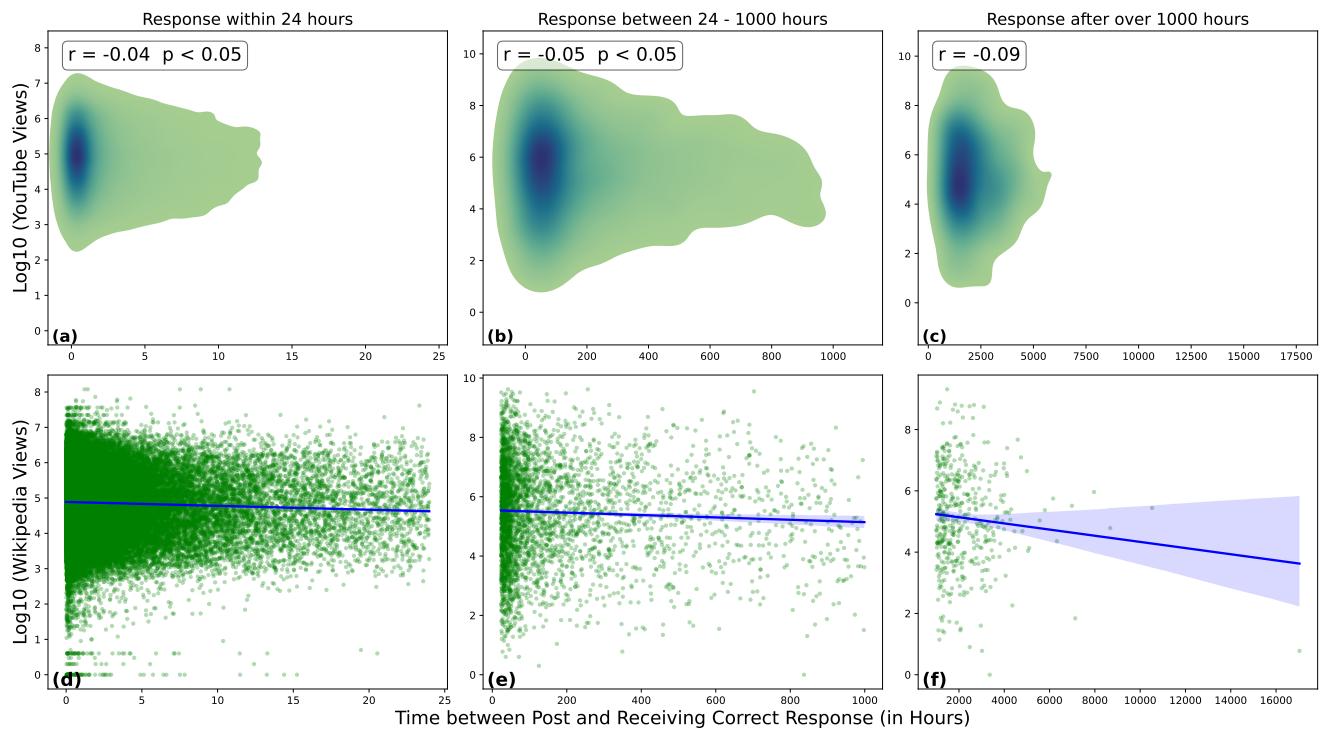


Figure 8. Zoomed-in correlation of Wikipedia-based popularity and response time on Reddit, to receive the correct answer.

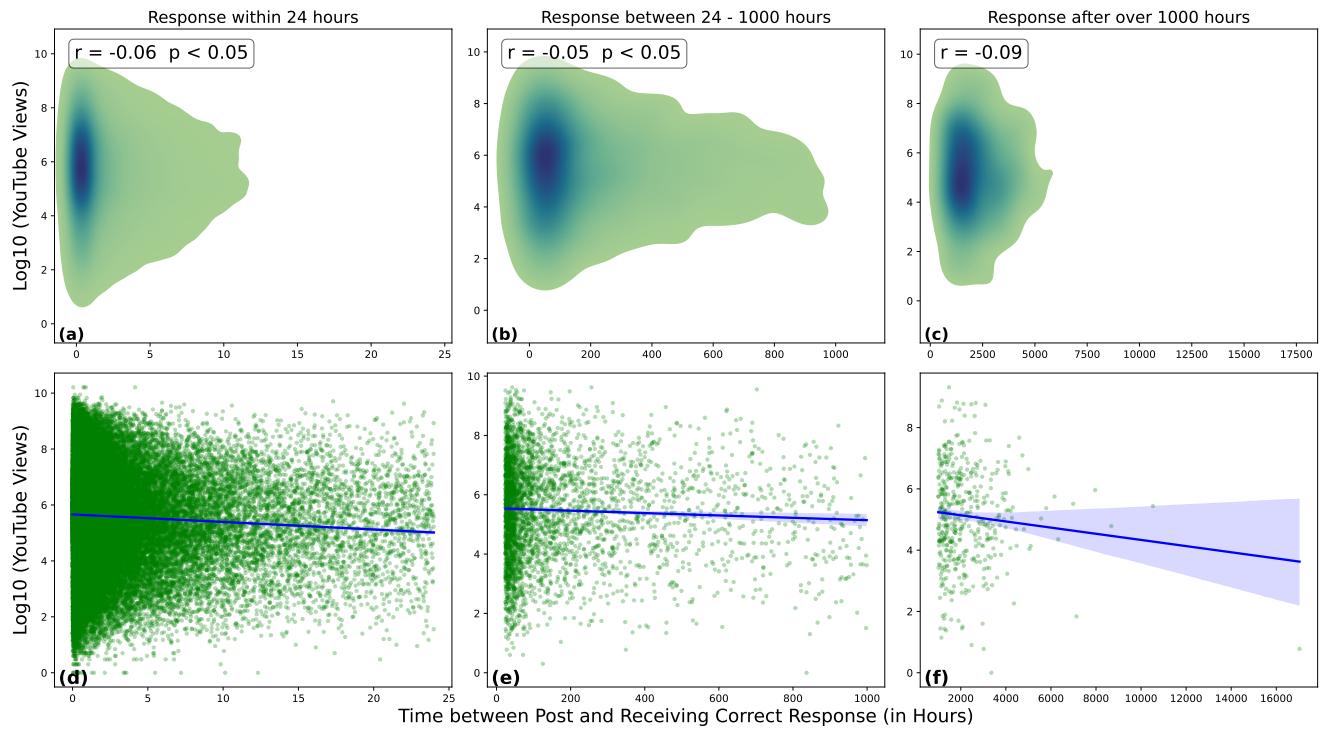


Figure 9. Zoomed-in correlation of YouTube-based popularity and response time on Reddit, to receive the correct answer.

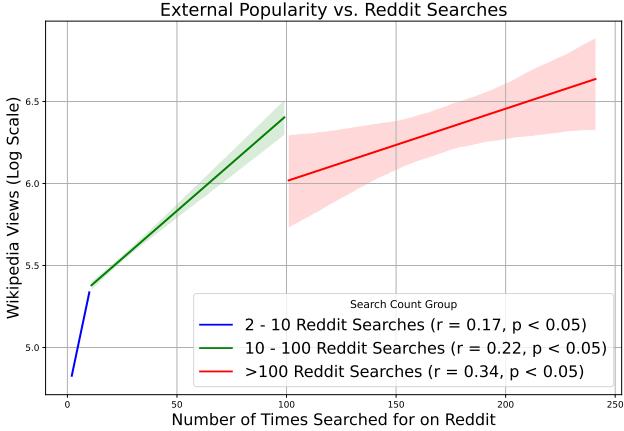


Figure 10. Condensed graph for correlation between Wikipedia page views (popularity) and number of searches on Reddit.

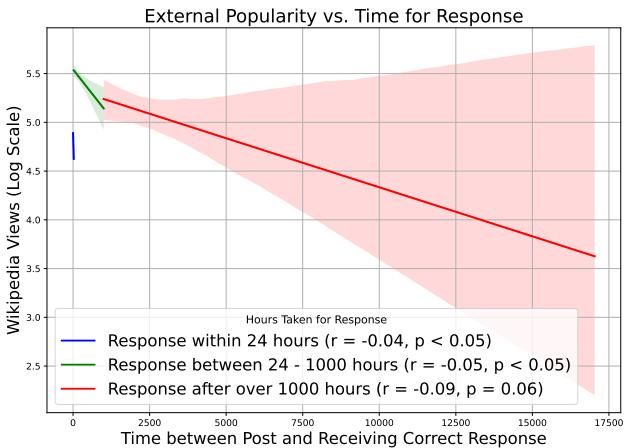


Figure 11. Condensed graph for correlation between Wikipedia page views (popularity) and time taken to receive the correct answer in response.

takes 8 hours. We also utilize DeepSpeed Zero and Flash Attention for training.

**Mining Hard Negatives.** Hard negative targets are chosen for each video-recall pair, during the training process for the retrieval task, based on how semantically similar the other recall signals are. For this, we first embed all ground truth recall statements using SBERT [43], and compute pairwise cosine similarity. For each sample, we then choose the hard negative recall statement,  $t^-$ , by randomly sampling from the top 50 most similar other recall statements.

Our goal in creating the hard negatives is to ensure that the embedding model learns to distinguish between samples where the memorability signals are the most similar to each other. In the text-based hard negative mining process, cap-

turing direct memorability signals is relatively easy, as we have access to the ground truth recall queries. On the contrary, using multimodal hard negatives is challenging, due to the lack of a precedent method that can embed videos (or find similarities between videos) based on memorability. Thus, we currently focus on using only text-based hard negatives.

**Task Instructions.** Here, we provide the different task instructions used, for the Recall Generation, Prompt Recall Ranking, and Retrieval task:

- **Instruction for Recall Generation:** “*You are given a detailed description of a video, including the audio transcript of the video, description of each scene in the video, and the text shown in each scene. Your task is to respond with what a person may say, when they are trying to remember the video. Precisely, if a person vaguely remembers the video, and is trying to retrieve a description of the video from their own memory, what are some possible things that they may say? Answer by considering all information about the given video: Audio Transcript: ...., OCR: .....*
- **Instruction for Prompt Recall Ranking:** “*You will be given multiple images, which are scenes from a video. The images are about some brand, depicting a brand advertisement. There are also several potential descriptions available for the sequence of images, which highlight what is most memorable from the video. Your task is, given 5 candidate descriptions for the images potential descriptions available for the sequence of images, choose the description which is the most fitting. You are required to answer strictly in a JSON format, providing the final answer as follows: answer: <Option n>*
- **Instruction for Retrieval:** “*You are given the scenes from an advertisement video, and a detailed description of the video, including description of each scenes in the video, audio transcript of the video, and title of the video. Your task is to respond with what a person may say, when they are trying to remember this advertisement video. Precisely, if a person vaguely remembers the video, and is trying to retrieve a description of the video from their own memory, what are some possible things that they may say? Answer by considering all information about the given video: Audio Transcript: ...., OCR: .....*

Note that the instruction for the retrieval task is similar to the recall generation task, and it directs the model to learn how to embed the given videos and recall queries to best capture signals of memorability.

## E. Additional Results for Recall Generation

Here, we present additional analysis and results for the recall generation task.

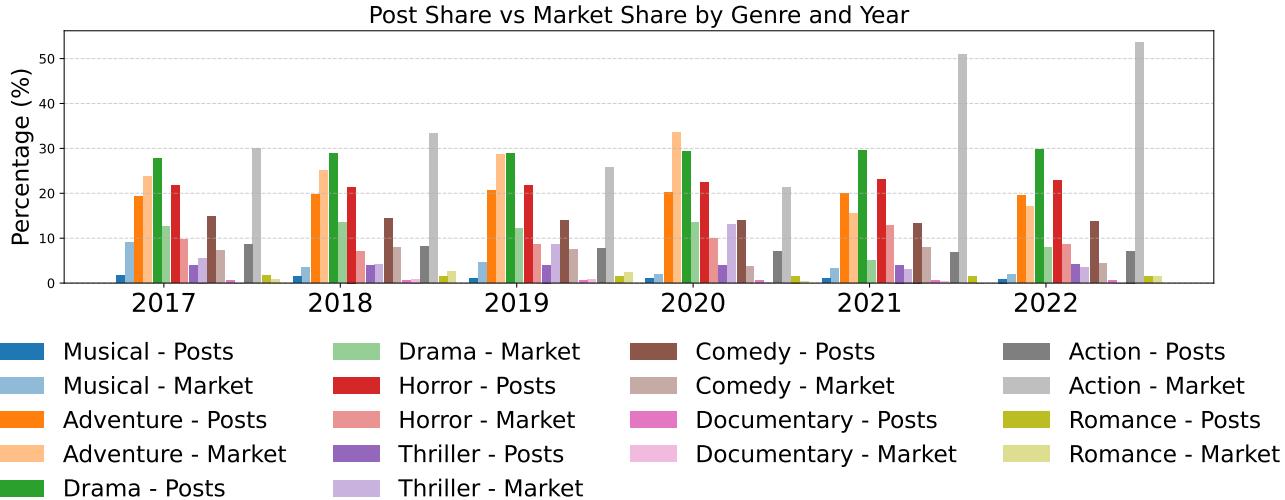


Figure 12. A detailed view comparing genre popularity based on external metrics - in this case, the average revenue earned by each genre within North America - with the corresponding popularity on Reddit ToT search platforms.

Model	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
InternVL-2.5 8B	0.21	0.196	0.216	0.059	0.148	0.82
InternVL 2.5 8B w/ TOT2MEM	<b>0.24</b>	<b>0.27</b>	<b>0.28</b>	<b>0.09</b>	<b>0.192</b>	<b>0.85</b>

Table 6. Results of fine-tuning another strong baseline using our dataset. This further supports the effectiveness of the dataset, and shows that the gains do not stem from a specific kind of backbone architecture.

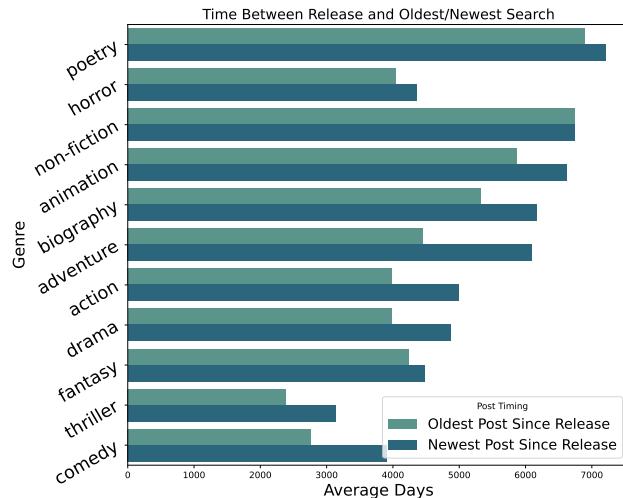


Figure 13. Average number of days since release that the oldest and latest posts are made for each content item, grouped by Genre. The creation date of the Wikipedia page for a given content item is used as the proxy for creation date.

**Fine-tuning other baselines:** As described in 4, our presented model TOT2MEM-RECALL is a version of

Qwen 2.5 VL 7B [2], fine-tuned on TOT2MEM. We use this setting to specifically demonstrate the effectiveness of our dataset, showing that without architectural changes, simply fine-tuning a baseline with our dataset can help learn a generalizable memorability signal. As an additional experiment, we also fine-tune InternVL 2.5 8B with TOT2MEM, and find similarly improved performance, as shown in Table 6.

**Controlling for Proper Noun Leakage in OCR, ASR:** Next, to understand whether the leakage of proper nouns through the automatically generated OCR and ASR provides an unfair advantage to the TOT2MEM-RECALL model at inference time, we perform an ablation study. We mask out all of the proper nouns in the ASR and OCR, simply replacing them with an empty string. We avoid using an explicit mask token (e.g., the commonly used [MASK]) to ensure that the model does not get further confused. We use the Spacy library <sup>10</sup> to remove proper nouns pertaining to the following: individual people names, names of organizations (e.g., companies, agencies, institutions, etc.), geopolitical entity names (e.g., country, city, state), non-political locations (e.g., mountains, rivers, etc.), creative

<sup>10</sup><https://spacy.io/>

Model	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
TOT2MEM-RECALL	0.242	0.293	0.304	0.152	0.251	0.85
TOT2MEM-RECALL w/ masked OCR, ASR	0.241	0.291	0.303	0.153	0.252	0.85

Table 7. Results of evaluating our model with test data where proper nouns are masked out from the generated OCR and ASR, to prevent leakage of information. The model performance remains relatively unchanged.

titles, named historical, cultural, or sports events, named commercial products, and names of manmade facilities or landmarks (e.g., Eiffel Tower, Heathrow Airport, etc.). We manually verify the masking process for 20 samples from the dataset and find that proper nouns are removed from both the transcript and the OCR. We present the ablated results in Table 7. The performance of the model remains virtually unchanged, confirming that the model’s predictions are not driven by name leakage but by broader descriptive signals. It is also worth noting that in some cases, with the proper nouns masked out with an empty string (“”), the audio transcript, in particular, becomes slightly noisy. Our results show that our model is also robust to such variations. In other words, fine-tuning with our dataset equips it with the capability to learn generalizable memorability signals.

**Ablating OCR and ASR:** We also provide two additional ablation experiments. We ablate both the presence of OCR and the automatically generated transcript in training data, one by one, and show the results on two of the chosen metrics in Table 8. We find that both the OCR and audio transcripts contribute to the performance of the trained models, with the contribution of ASR being slightly more significant. We hypothesize that there may exist an intuitive reason for this. For videos where the visual content and the semantic meaning (or message) are disparate, the audio transcript provides a bridge between the two. It may also be significantly useful for the model to track temporal changes and relate them to the temporally changing scenes. Particularly, given that we only provide the model with sampled keyframes (limited to 30 scenes per video), the audio transcript becomes the only source of complete information about the video. However, even without either the OCR or the audio transcript, the model is capable of learning some memorability signals from the visual information, leading to improvements over its zero-shot version.

## F. Qualitative Examples

In this section, we provide qualitative examples of responses generated by the models fine-tuned on our dataset, TOT2MEM-RECALL and TOT2MEM-RETRIEVAL for the recall generation and ToT retrieval tasks, respectively.

**Examples for Recall Generation.** Figures 14 to 19 show the qualitative examples for high-scoring and low-scoring

generated recall samples. We use the BERTScore ratings for each generation to retrieve the most highly scored, and the lowest scored examples. BERTScore is chosen to provide the qualitative examples, due to its higher semantic matching capability compared to the other static metrics, like BLEU or ROUGE scores. We show 3 scene examples from each video, and omit directly providing the YouTube video ID or link, to preserve privacy for the video uploaders and channels.

Figures 14 to 16 show the top-rated generations by our model. Interestingly enough, in the second example (Fig. 15, the ground truth query refers to the video as a “Japanese boy and an English speaking girl”, while our model predicts that the video was a “Korean music video”. We hypothesize that the OCR is useful in this case, as the subtitles shown in Korean are picked up by the model as relevant memorability signals. Similarly, in the third example 16, our model is capable of capturing the temporal change of actions in the video, as it generates, “At one point the conductor starts dancing around...”. It can also be noted that the model, in alignment with human-generated recall, adds a sentence expressing confusion.

On the other hand, Figures 17 to 19 show the generations with some of the lowest scores. We show here 3 different cases of failures:

- **Gap Between Visual Features and Semantics:** We find, as shown in Fig. 17, the generation from our model is highly aligned with the visual content shown in the video, while the ground truth recall query talks more about the semantic content or an abstract summary of the video. This shows that the trained model develops high visual fidelity, and potentially develops a strong visual branch through the training process, different from previous work [21], where the model needed explicit textual verbalizations of the scenes to perform adequately well in predicting memorability scores. This also highlights the challenging and nuanced nature of the task, and can inform valuable future work, where the focus can be shifted to building better semantic or abstract representations of videos. This would become especially useful for cases where the visual content may not be well-aligned or entirely coherent with the semantic or underlying message.
- **Repetitions:** As shown in the second example (Fig. 18, our fine-tuned version of Qwen 2.5 VL, in some cases, falls into the well-known pitfall of repeating tokens. Fu-

Model	BLEU	ROUGE-1	BERTScore
TOT2MEM-RECALL	0.242	0.304	0.85
TOT2MEM-RECALL - OCR	0.23	0.26	0.84
TOT2MEM-RECALL - ASR	0.22	0.23	0.82

Table 8. Comparison of our model with ablated training setups. The “- OCR” denotes training without the per-scene OCR texts, and “- ASR” denotes training without the audio transcript for each video.

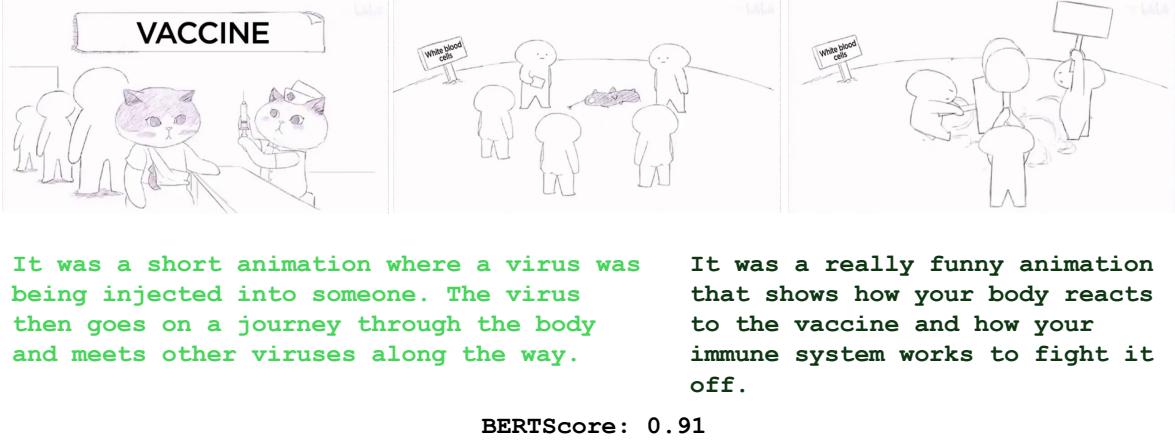


Figure 14. Example 1 of a highly scored recall generated by our model TOT2MEM-RECALL. In this example, the text on the **left shows the prediction**, while the **text on the right shows the ground truth recall**. The corresponding BLEU Score for the generation is 0.46, and ROUGE-1 score is 0.39.

ture work can explore in depth the specific origins of repetition in a memorability-related task [53], or general-purpose solutions such as employing nucleus sampling [23]. In fact, we experiment with using nucleus sampling specifically and find that, although the overall performance does not degrade, the issue of tokens being repeated remains.

- **Presence of Links:** As shown in the final error example (Fig. 19), we find a small number of samples where the output is only a link. As previously discussed in Sections 3 and A, given our automated mechanism of data filtering, for some of the queries, some link content may remain retained. Future work could focus on penalizing the generation of specific kinds of hallucinated text, especially links that do not exist. This could also inform important future work on further exploring what types of hallucinations are possible in the realm of memorability-related tasks.

**Examples for Text-to-Video ToT Retrieval.** We further provide similar examples for the multimodal ToT retrieval task. Specifically, we show two types of examples: the first, where the correct video is retrieved within the top-5 items, and the second, where the correct video is not retrieved even within the top-100 items. For each category, we show the

original recall query, scenes from the original correct video. For the first category, we also provide examples of one other video which are retrieved in the top-5 items (highly similar to the original video). For the second category, we provide examples of the top-most (incorrect) video. The examples are shown in Figures 20 and 21.

In the correctly retrieved example (Fig. 20), the two videos are similar in quality and style. Both are videos of games, and the recall for both gives the description of natural elements. In the incorrectly retrieved example (Fig. 21), the common elements between the original and incorrect videos seem to be the content type (song), and mentions of a particular accent, or a particular group of people (“Australian”, “European”). Future studies on what is the most optimal information to retain from a recall signal may also be a promising direction.

## G. Additional Details about Human Evaluation

As described in Section 4, we use a curated experiment involving human data to evaluate the robustness of the recall-generator model, trained on our unsupervised dataset. We collected the data from 163 participants at an academic institution in India. Participation was primarily voluntary, but



I think it was a Korean music video. It was a girl and boy singing together.

I believe it was a Japanese boy and an English speaking girl. It was made by one of those Facebook video companies with a studio set up.

BERTScore: 0.93

Figure 15. Example 2 of a highly scored recall generated by our model TOT2MEM-RECALL. In this example, the text on the **left shows the prediction**, while the **text on the right shows the ground truth recall**. The corresponding BLEU Score is 0.41, and ROUGE-1 score is 0.62.



It was a classical orchestra playing a piece. At one point the conductor starts dancing around the stage and the audience starts clapping. I think it's a famous piece but I can't find it.

It starts with the conductor playing a normal symphony then I think he turns around and gets the audience to clap to the beat of the song. I think it's a fairly short video like 1 min.

BERTScore: 0.91

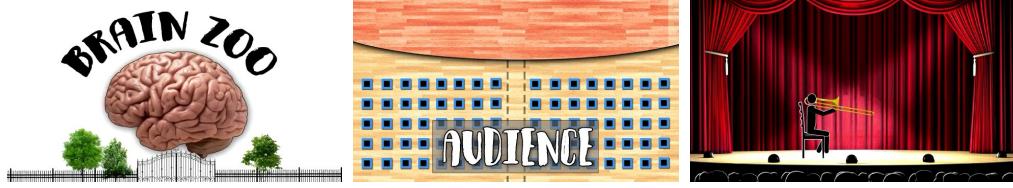
Figure 16. Example 3 of a highly scored recall generated by our model TOT2MEM-RECALL. In this example, the text on the **left shows the prediction**, while the **text on the right shows the ground truth recall**. The corresponding BLEU Score is 0.61, and the ROUGE-1 score is 0.56.

students were also provided with the option to get optional course credit, or other freebies. The data collection process was reviewed and approved by an IRB within the academic institution. The students were required to sign the IRB consent approval, which was displayed clearly before the data collection. The consent approval information provided details on the kind of data being collected, how the data would be used, and the time commitment required for the study. The participants were provided with the aggregate statistics of the study as a debrief after completing the study. Participants could either take the study in a take-home or in-person format. For take-home participants, similar to the original protocol for the collection of the LAMBDA dataset [21], three emails were sent, and if they did not respond, their data was discarded from the experiment.

The participants were primarily graduate and undergrad-

uate students. The participants are bilingual and speak a variety of languages, including English. The age range of the participants is from 18 to 35 years, and there was again a roughly 30 - 70 % distribution of female and male participants. We did not record additional gender identification information, although participation from all genders was encouraged.

Participants were instructed to log into a controlled environment hosted by us and recreate memorable ad scenes using generative models (Stable Diffusion v2.1) [45]. The generated images and prompt logs were stored in a submission archive organized by brand and attempt number. The detailed questionnaire for this reconstruction task is included in §G.2.

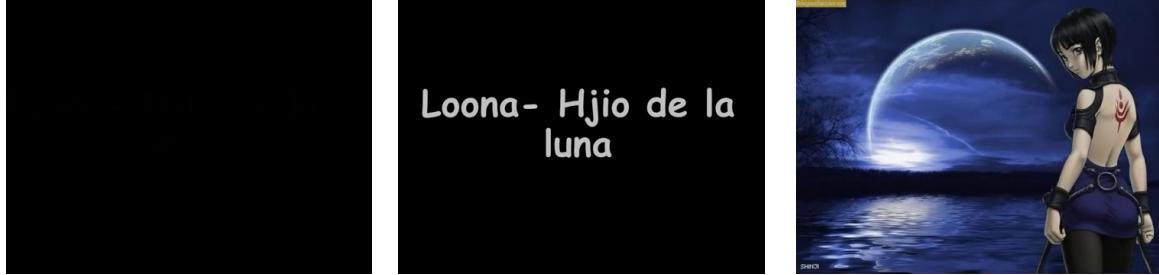


I remember seeing a video on YouTube a few years ago about the brain. It was animated and had a narrator. The narrator explained that the brain is like a zoo and that there are different animals in the brain. He also said that the brain is like a concert hall and that the audience is the part of the brain that is aware of the world around it. The narrator also said that the orchestra is the part of the brain that is responsible for creating the music that we hear in our heads. The narrator also said that the brain is like a computer and that the brain is like a computer.

They said something about the good feelings being a good concert, with the performers playing pleasantly, and in contrast the bad feelings are some performers who are playing bad and unpleasantly.', 'I also remember them stating that when that happens the audience joins in and starts screaming in despair, and thus the whole audience becomes part of the horrible performance.

BERTScore: 0.77

Figure 17. Example 3 of a highly scored recall generated by our model TOT2MEM-RECALL. In this example, the text on the **left** shows the prediction, while the **text on the right** shows the ground truth recall. The corresponding BLEU Score is 0.10, and ROUGE-1 score is 0.11.



I think it was a Spanish song. It had a female singer and the chorus went something like "Luna, luna, luna" or "luna, luna, luna, luna.". I think it was a Spanish song. It had a female singer and the chorus went something like "Luna, luna, luna" or "luna, luna, luna, luna".

I can't tell whether its from 60's 70's 80's or 90's. It was quite popular and known I guess. Could be folk music. All I can say is that there was a female vocalist and it was kinda a faerie/night/mystical/fantasy ambiente (and perhaps slightly melancholic) and maybe middle age themed. I assume the singer has dark hair, but not sure though.

BERTScore: 0.78

Figure 18. Example 3 of a highly scored recall generated by our model TOT2MEM-RECALL. In this example, the text on the **left** shows the prediction, while the **text on the right** shows the ground truth recall. The corresponding BLEU Score is 0.10, and ROUGE-1 score is 0.13.

### G.1. Performance on Prompt Recall Ranking Task

As seen in 4, Table 4, there is a significant performance gap between the baseline models InternVL and Qwen 2 VL. From a qualitative analysis of responses, we find that InternVL struggles particularly with producing output in the

required format, where it is required to choose between multiple options. On the other hand, Qwen 2 VL excels at providing output in the correct format. The performance shown by our model, TOT2MEM-RECALL, becomes particularly noteworthy here. Although it is a model fine-tuned



<https://youtu.be/0QcZvqYJrUg?t=1m1s>

BERTScore: 0.77

I suspect it's a pretty new and recent song. It was a female artist. She sang in a bit of a mumbly way, or at least not very loudly, like maybe she had a careful or sad vibe to her. The song wasn't upbeat at all; it was quite slow. The style was kinda like how Billie Eilish sometimes sings (her song 'everything I wanted' is a great example of what I mean) or like Eiza Murphy, if you're more familiar with that name. In terms of lyrics I couldn't quite catch any of them for sure cuz of the noise in the store, but I'm pretty sure the lines were all rather short and one of them started with her kinda singing / SIGHING the words: 'I get so mad, ain't it sad'. At some point she also says 'difficult'

Figure 19. Example 3 of a highly scored recall generated by our model TOT2MEM-RECALL. In this example, the text on the left shows the prediction, while the text on the right shows the ground truth recall. The corresponding BLEU Score is 0.00004, and ROUGE-1 is 0.09.

for a completely different task, it is still able to output responses in the correct format, and does so more accurately than the other baselines. We conjecture that the low-rank fine-tuning may have contributed to this, where the model has an improved ability to recognize memorability-related signals, while at the same time being able to respond in accordance with instructions.

## G.2. Scene Reconstruction Questionnaire

This section contains the additional questions we asked as part of the study, where participants were asked to recreate memorable advertisement scenes using generative models. The reconstructed outputs, prompts, and questionnaire responses were collected for each brand remembered.

### G.2.1. Scene Description and Reconstruction

We hosted a website to allow students to simultaneously create scenes, we hosted SD 2.1 on 4 nodes of A100 GPUs. An exemplar scene creation on the platform is shown in figure 22

1. For the {brand} ad, I remember seeing the following:
  - (Write scene descriptions, feel free to include any scenes, music, characters, emotions, or objects you remember seeing)
2. Please go to the website and recreate a scene that you remember from the {brand} ad.
  - Fill your prompts and outputs in prompts.txt and

store them in folders #1, #2, #3, #4, #5.

- You may attempt up to 5 reconstructions per remembered scene.

### G.2.2. Audio Recall (to be filled for each reconstructed brand ad)

1. For the {brand} ad(s), what type of audio did you hear? (Select all that apply)
  - Narration
  - Background Music
  - Silent
  - Don't Remember

### G.2.3. Product Familiarity (to be filled after reconstruction)

1. How many times in the last 1 year have you used the product shown in the {brand} ad(s)?
  - 0
  - 1–10
  - 10+
2. Have you ever used {brand} before?
  - Yes
  - No



Figure 20. This is an example of a retrieval output, where text-to-video retrieval is performed by our model TOT2MEM-RETRIEVAL. In this case, the correct video is retrieved within the top-5 elements. We show the correct video along with another similar video, which is retrieved within the top-5 items.

Original Query

I'm looking for a song (pre 2010) that I remember hearing on the radio/tv with my parents multiple times. I know it's fairly famous, and I seem to remember it having a particular affinity with Australian sporting events/Australia generally. I might be making that up, as I'm from the UK, but I seem to distinctly remember a crowd of aussies singing it. I wonder if it might have been sweet caroline but it doesn't feel right. Any ideas (I know it's very vague).

Original Video



Retrieval Result

Not Retrieved Correctly within Top-100.

Similar Video



Associated Recall

I think it'd technically be classic rock or just rock in general. It has a title like "Dum Dum Dum" or something with short repeating words. The album art is either bright yellow or pink. The singer has a bit of an irish or european accent.

Retrieval Result

Retrieved Incorrectly within Top-100 at Position 2.

Figure 21. An example of a retrieval output, where the correct video is not retrieved within the top-100 items. In this case, we show the correct video corresponding to the original recall query, which is not retrieved in top-100, and another video, which is incorrectly retrieved within the top-100 items.

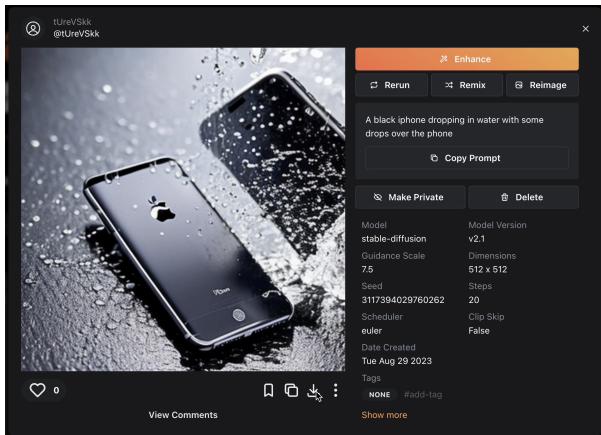


Figure 22. Image generation platform used by students, the prompt can be written here and they can retry variations upto five times.