Acknowledgments

S. Bhattacharyya and J. Z. Wang have been supported by the U.S. National Science Foundation (NSF) under Grant No. 2234195. They used cluster computers at the National Center for Supercomputing Applications and the Pittsburgh Supercomputing Center through an allocation from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF Award Nos. 2138259, 2138286, 2138307, 2137603, and 2138296. The authors are grateful to the reviewers for their thoughtful and constructive comments.

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 Conference on Human Information Interaction and Retrieval*, pages 5–14, 2021. 3, 8

[2] Shuai Bai, Keqin 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, 16

[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. 5

[4] Amin Beheshti, Shahpar Yakhchi, Salman Mousaeirad, Seyed Mohssen 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 reranking. 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, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan 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. 2

[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. *arXiv preprint arXiv:2507.05595*, 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 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. 5

[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 *Proceedings of the QPP++ Workshop at the 45th European Conference on Information Retrieval (ECIR)*, pages 13–19, 2023. 3

[18] Francis Galton. *Vox populi*. Nature Publishing Group UK London, 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, 14

[21] SI Harini, Somesh Singh, Yaman K Singla, Aanisha Bhattacharyya, 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, 2, 4, 6, 7, 20, 22

[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. 20

[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, 15

[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 *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 145–152. IEEE, 2011. 1, 2

[27] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representation*, 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-Héléne Demarty, Graham Healy, Bogdan Ionescu, and Alan F Smeaton. An annotated video dataset for computing video memorability. *Data in Brief*, 39:107671, 2021. 2

[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. 5

[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. 2, 4

[39] Dale Owen. Designing with memory in mind. *UX Collective*, 2019. Accessed: 2025-12-02. 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. 5

[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 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019. 15

[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, 23

[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, Wenhai 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. 20

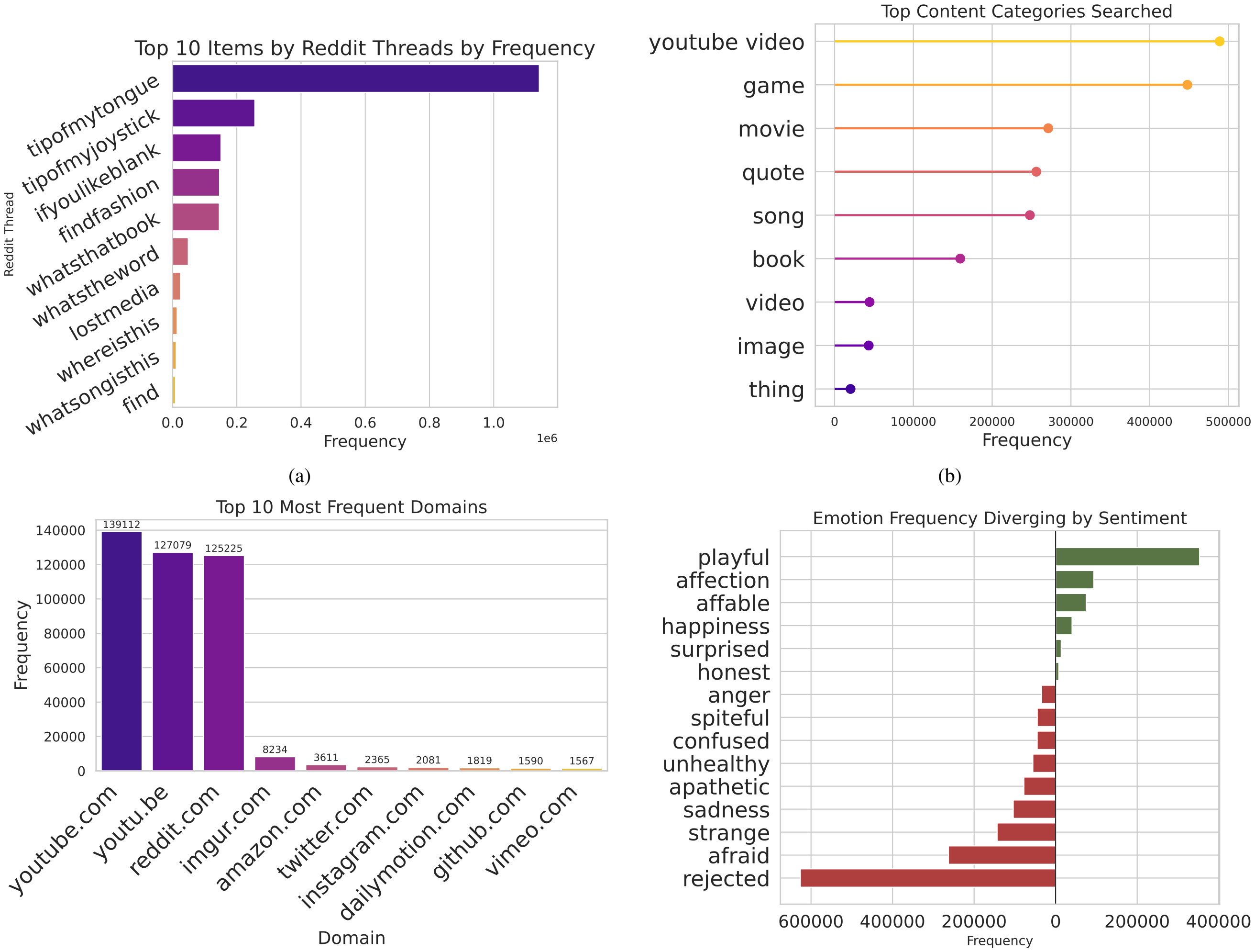
[54] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024. 7

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 subsection, we provide some additional statistics of the entire TOT2MEM dataset collected from Reddit. First, 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)

(b)

(c)

(d)

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].

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.

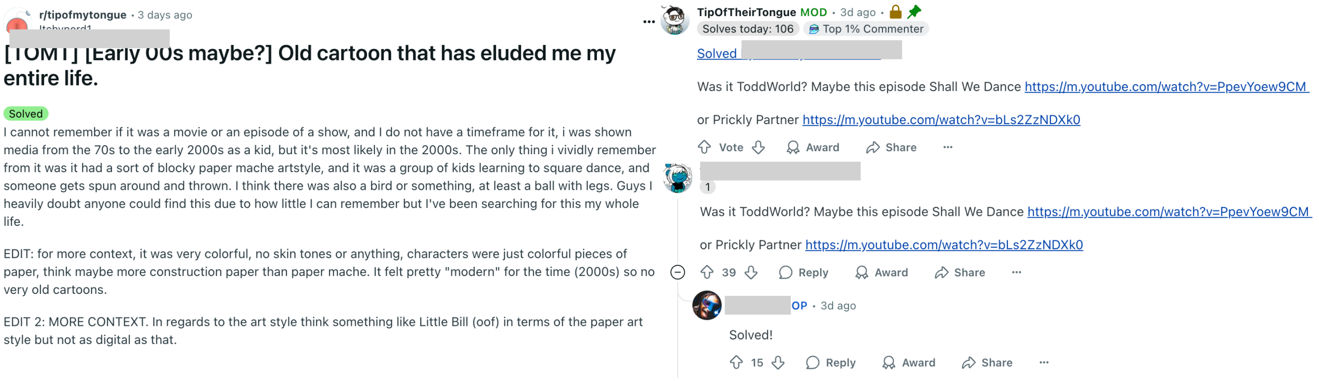
C. Connecting with External Factors

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 [9].

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.

[9] https://www.the-numbers.com/



r/tipofmytongue • 3 days ago

[redacted]

[TOMT] [Early 00s maybe?] Old cartoon that has eluded me my entire life.

Solved

I cannot remember if it was a movie or an episode of a show, and I do not have a timeframe for it, i was shown media from the 70s to the early 2000s as a kid, but it's most likely in the 2000s. The only thing i vividly remember from it was it had a sort of blocky paper mache artstyle, and it was a group of kids learning to square dance, and someone gets spun around and thrown. I think there was also a bird or something, at least a ball with legs. Guys I heavily doubt anyone could find this due to how little I can remember but I've been searching for this my whole life.

EDIT: for more context, it was very colorful, no skin tones or anything, characters were just colorful pieces of paper, think maybe more construction paper than paper mache. It felt pretty "modern" for the time (2000s) so no very old cartoons.

EDIT 2: MORE CONTEXT. In regards to the art style think something like Little Bill (oof) in terms of the paper art style but not as digital as that.

TipOfTheirTongue MOD • 3d ago •

Solves today: 106 Top 1% Commenter

Solved [redacted]

Was it ToddWorld? Maybe this episode Shall We Dance https://m.youtube.com/watch?v=PpevYoew9CM

or Prickly Partner https://m.youtube.com/watch?v=bLs2ZzNDXk0

[redacted]

1

Was it ToddWorld? Maybe this episode Shall We Dance https://m.youtube.com/watch?v=PpevYoew9CM

or Prickly Partner https://m.youtube.com/watch?v=bLs2ZzNDXk0

[redacted] OP • 3d ago

Solved!

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!").

We include several additional analysis results, complementing the genre-specific and popularity-related analysis, and include them in Figs. 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 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, capturing 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.

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 [10] 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 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. Qualitaive 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. Figs. 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.

Figs. 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.

[10] https://spacy.io/

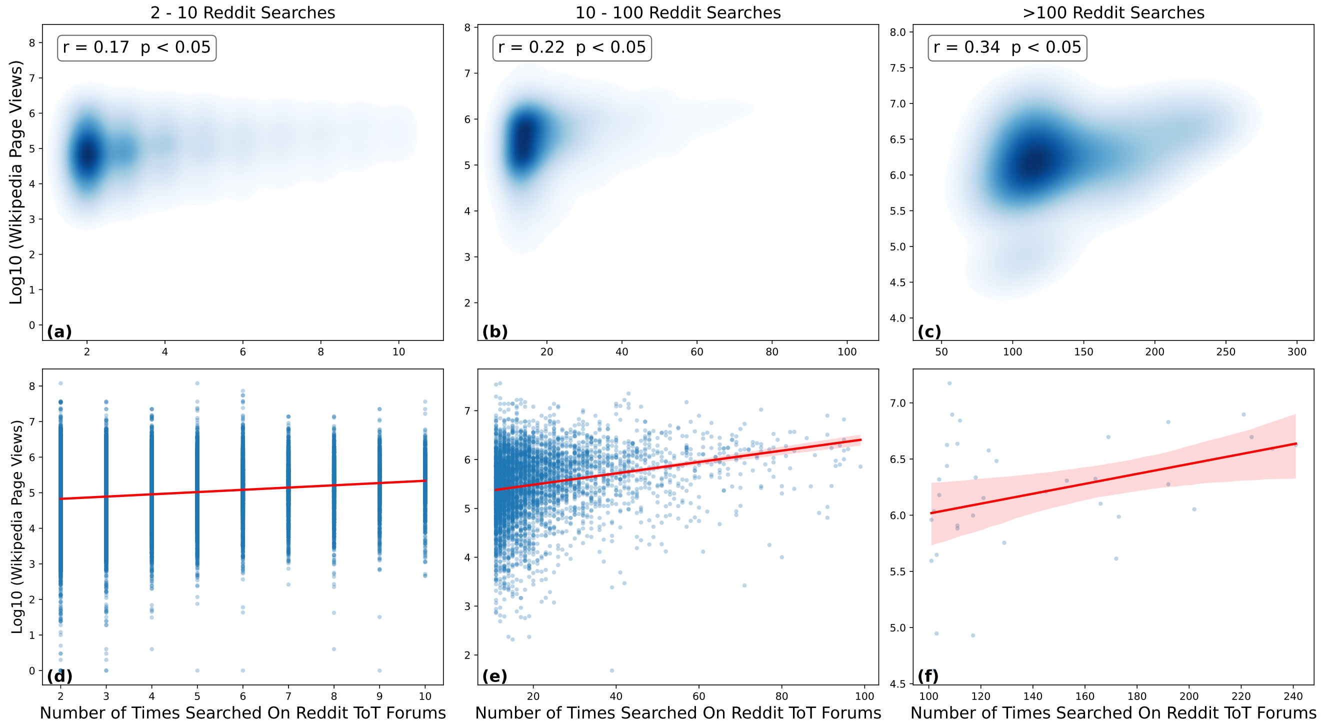


Figure 6. Correlation of Wikipedia-based popularity and number of searches on Reddit, for any given content item, shown in different groups, based on the number of searches made.

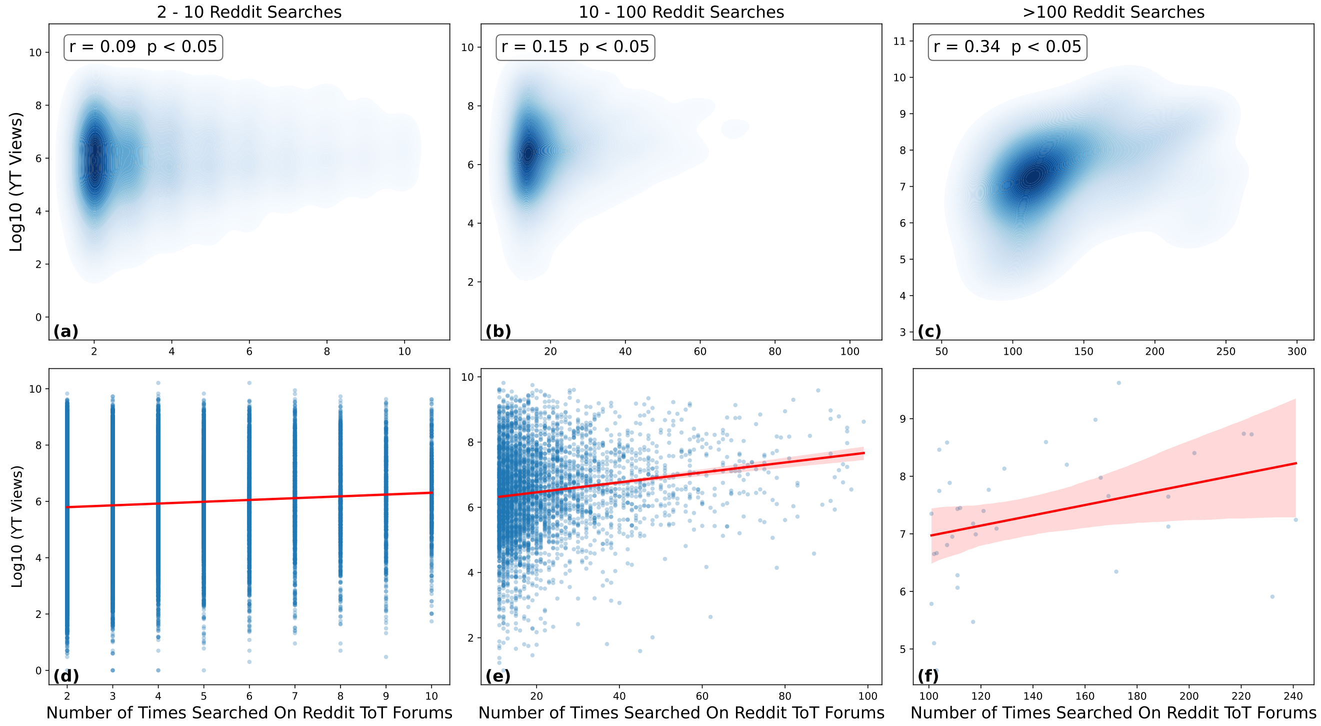


Figure 7. Correlation of YouTube-based popularity and number of searches on Reddit, for any given content item, shown in different groups, based on the number of searches made.

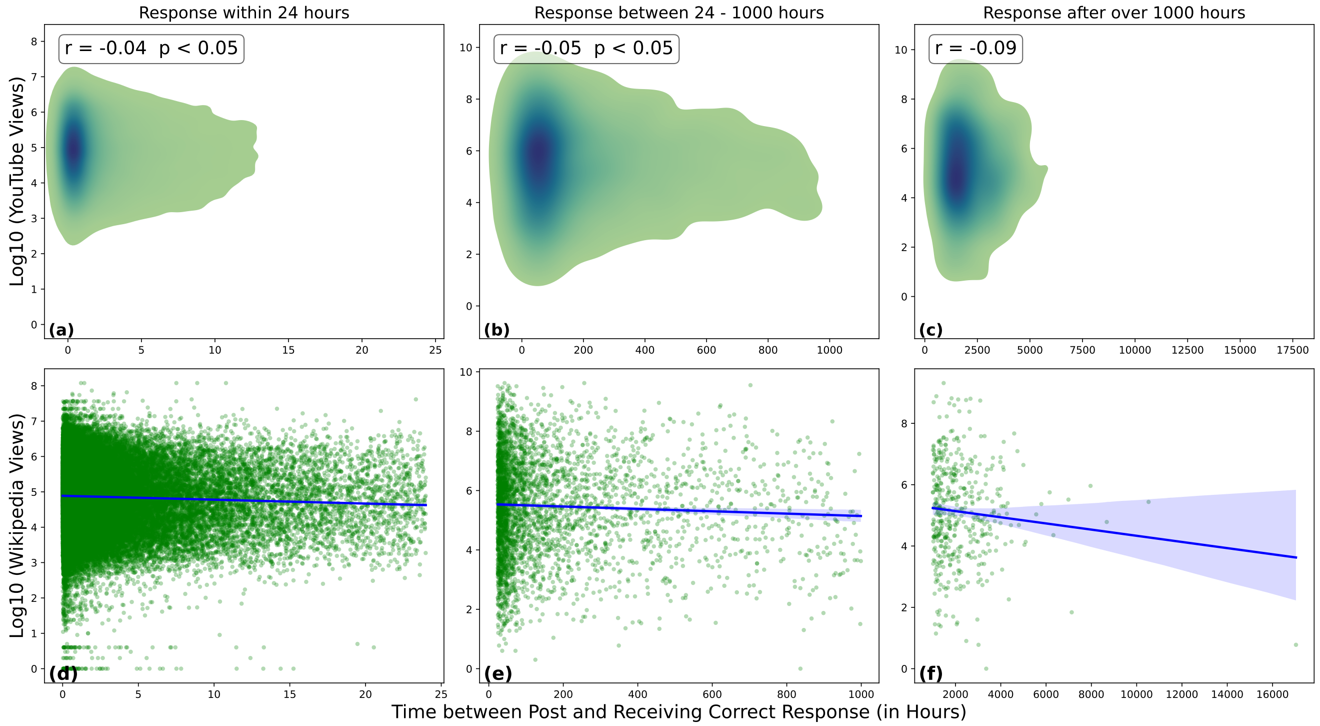


Figure 8. Zoomed-in correlation of Wikipedia-based popularity and response time on Reddit, to receive the correct answer.

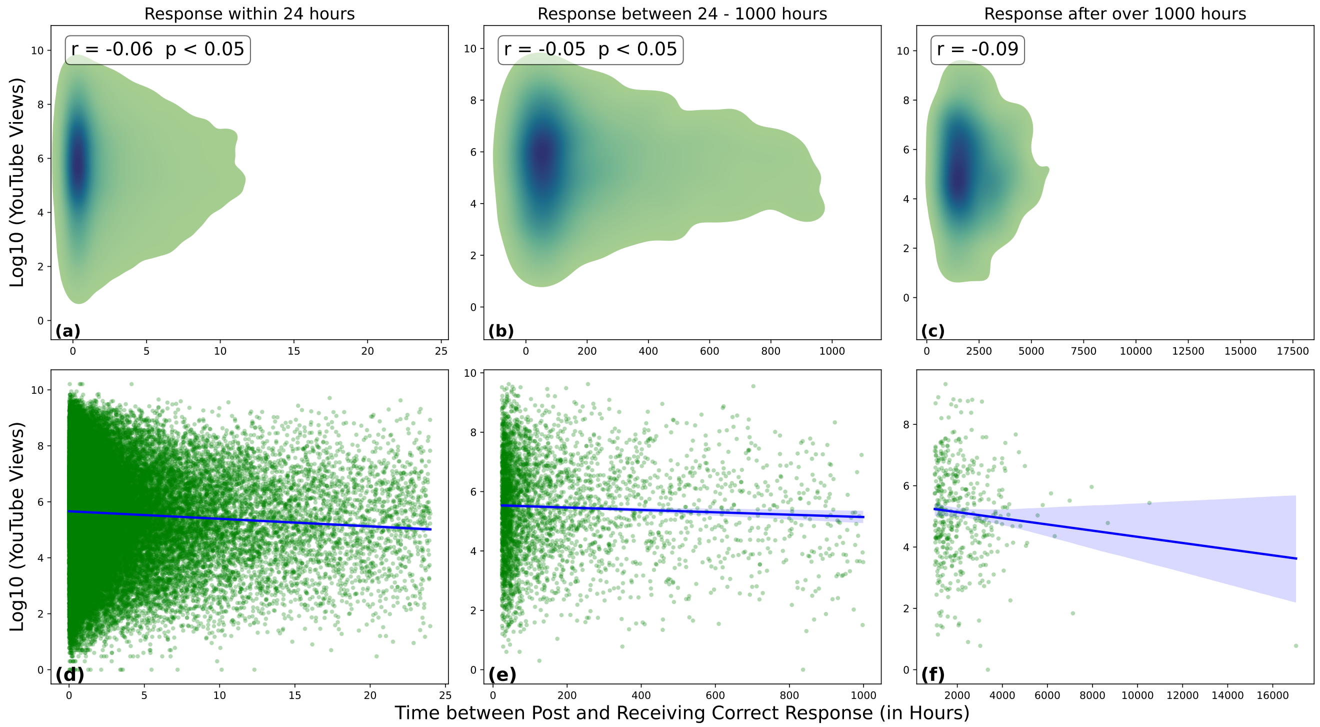


Figure 9. Zoomed-in correlation of YouTube-based popularity and response time on Reddit, to receive the correct answer.

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	**0.24**	**0.27**	**0.28**	**0.09**	**0.192**	**0.85**

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.

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.

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.

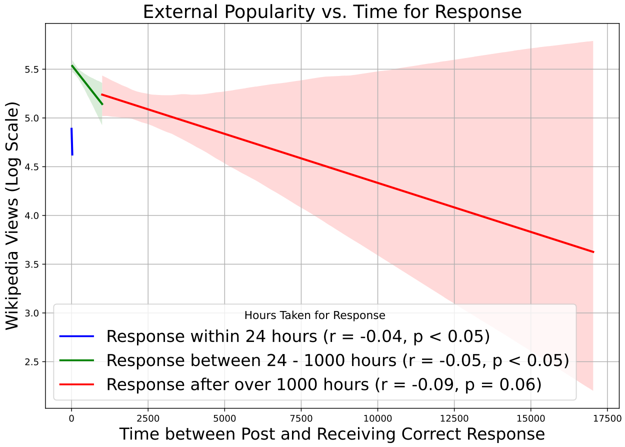


Figure 11. Condensed graph for correlation between Wikipedia page views (popularity) and time taken to receive the correct answer in response.

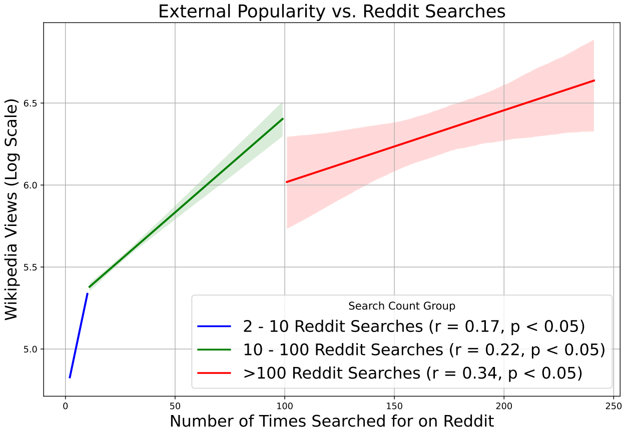


Figure 10. Condensed graph for correlation between Wikipedia page views (popularity) and number of searches on Reddit.

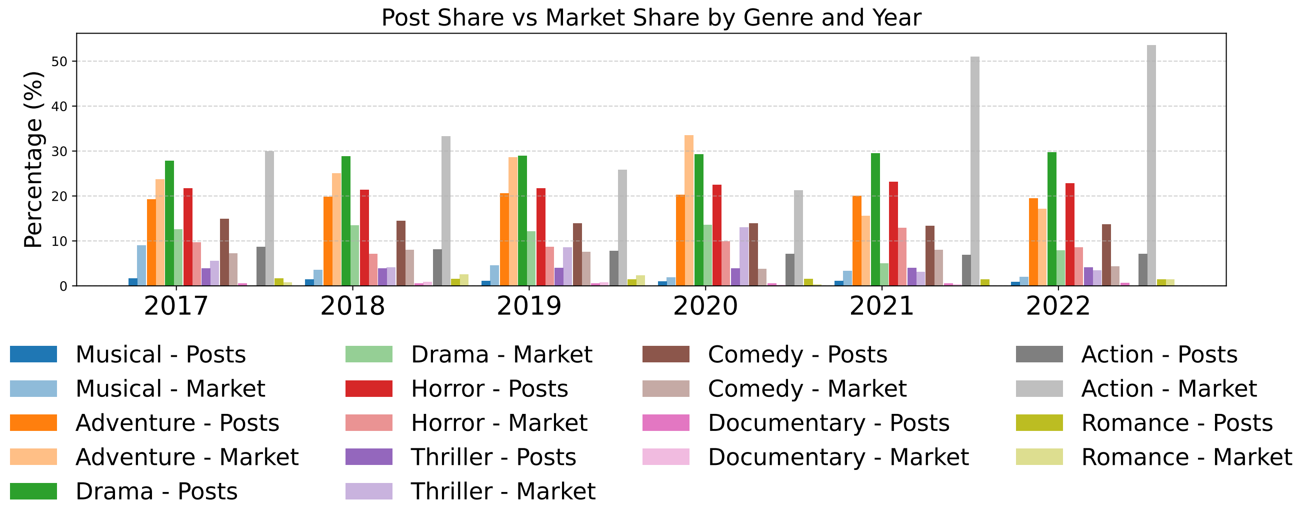


Figure 12. A detailed view comparing genre popularity based on external metrics - in this case, the average revenue earned by each genre within North America - with the corresponding popularity on Reddit ToT search platforms.

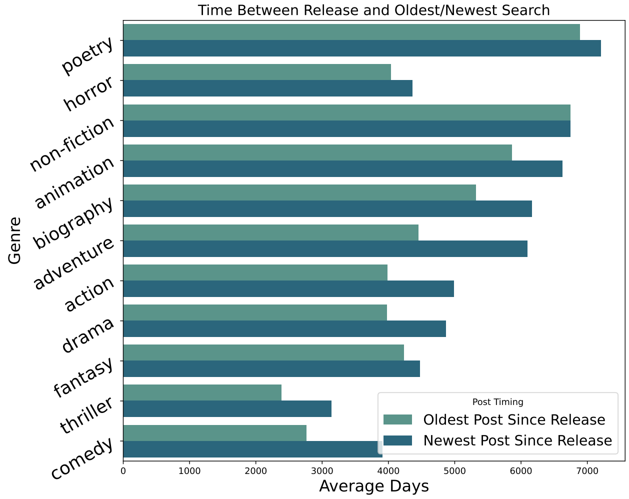
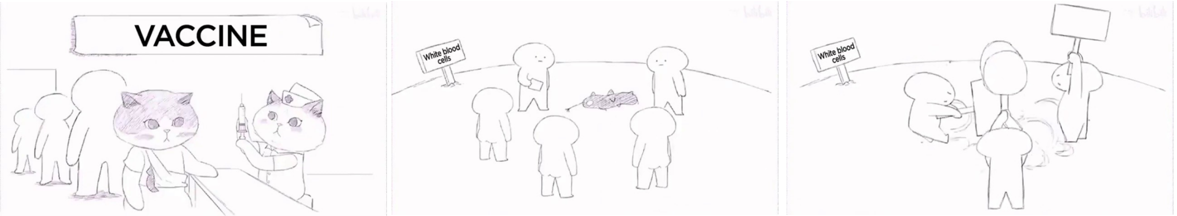


Figure 13. Average number of days since release that the oldest and latest posts are made for each content item, grouped by Genre. The creation date of the Wikipedia page for a given content item is used as the proxy for creation date.

On the other hand, Figs. 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. Future 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.

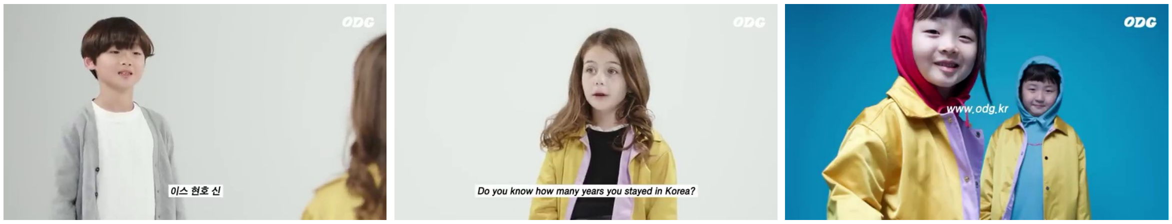


It was a short animation where a virus was being injected into someone. The virus then goes on a journey through the body and meets other viruses along the way.

It was a really funny animation that shows how your body reacts to the vaccine and how your immune system works to fight it off.

BERTScore: 0.91

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.



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.

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 Figs. 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

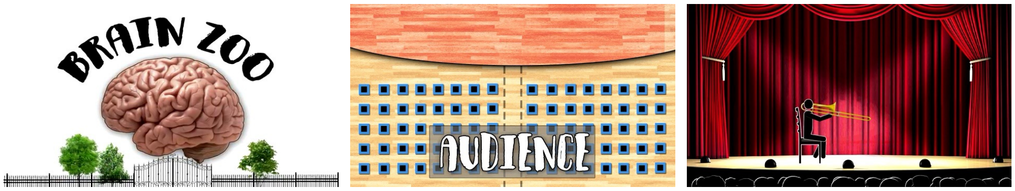
As described in Section 4, we use a curated experiment involving human data to evaluate the robustness of the recall-generator model, trained on our unsupervised dataset. We collected the data from 163 participants at an academic institution in India. Participation was primarily voluntary, but students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

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.



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.

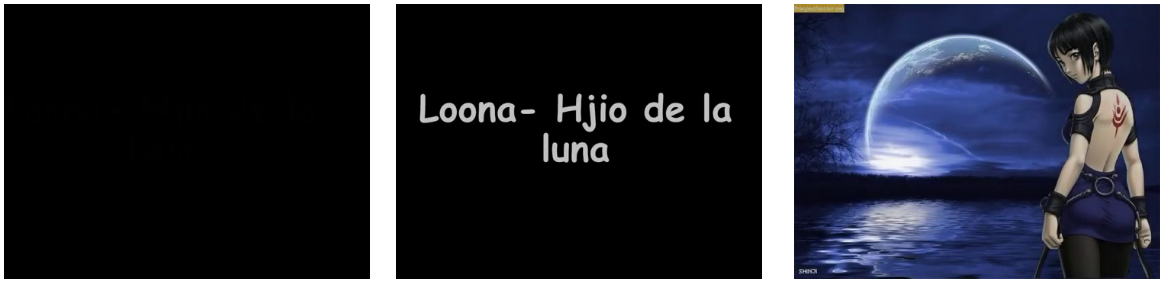
BERTScore: 0.77

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.

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 undergraduate 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

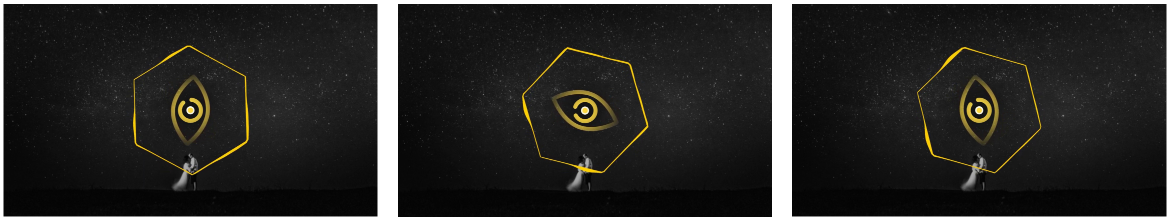


I think it was a Spanish song. It had a female singer and the chorus went something like "Luna, luna, luna" or "luna, luna, luna, luna.”. I think it was a Spanish song. It had a female singer and the chorus went something like "Luna, luna, luna" or "luna, luna, luna, luna”.

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.



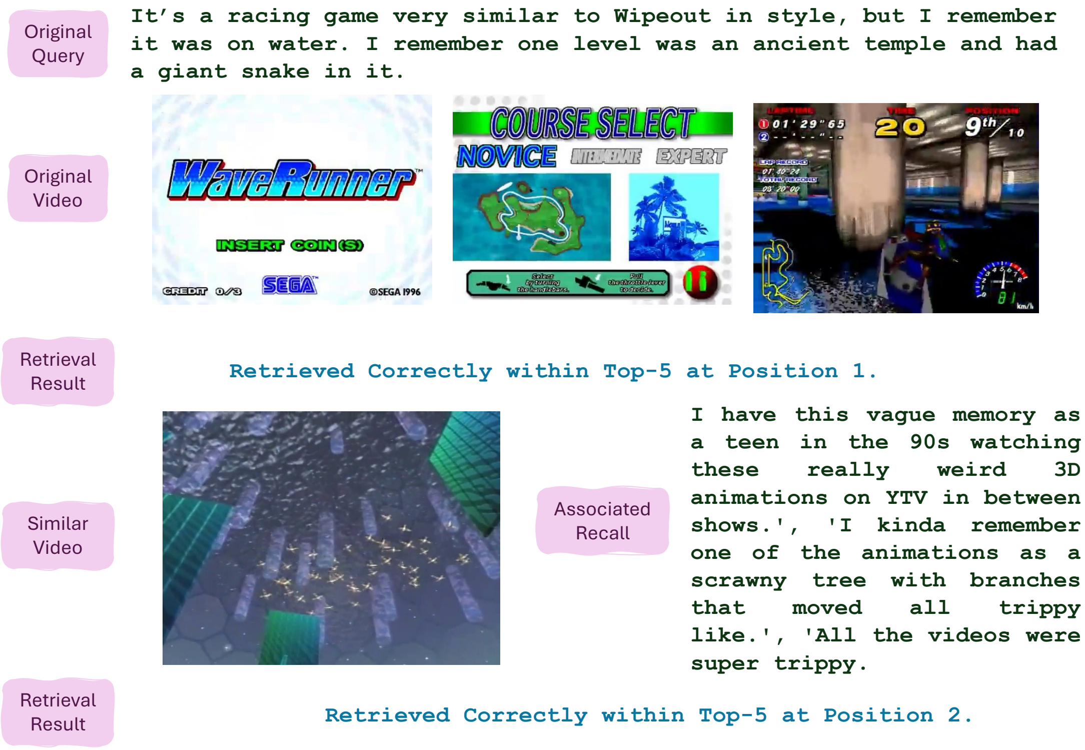
https://youtu.be/0QcZvqYJrUg?t=1m1s

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.

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.



It’s a racing game very similar to Wipeout in style, but I remember it was on water. I remember one level was an ancient temple and had a giant snake in it.

Retrieved Correctly within Top-5 at Position 1.

I have this vague memory as a teen in the 90s watching these really weird 3D animations on YTV in between shows.', 'I kinda remember one of the animations as a scrawny tree with branches that moved all trippy like.', 'All the videos were super trippy.

Retrieved Correctly within Top-5 at Position 2.

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.

The detailed questionnaire for this reconstruction task is included in §G.2.

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 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

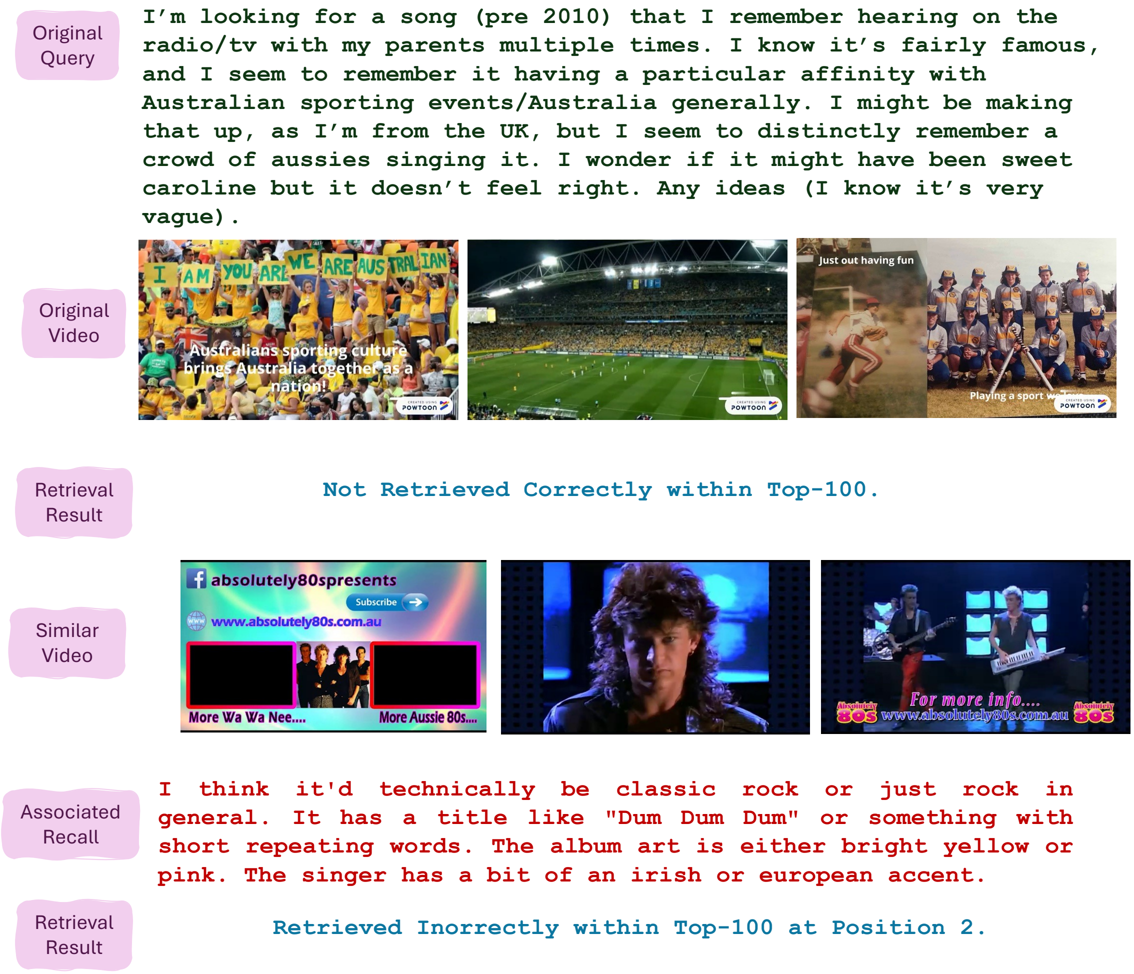
1. For the {brand} ad, I remember seeing the following:
   - (Write scene descriptions, feel free to include any scenes, music, characters, emotions, or objects you remember seeing)
2. Please go to the website and recreate a scene that you remember from the {brand} ad.
   - Fill your prompts and outputs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

   a. Narration

   b. Background Music



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).

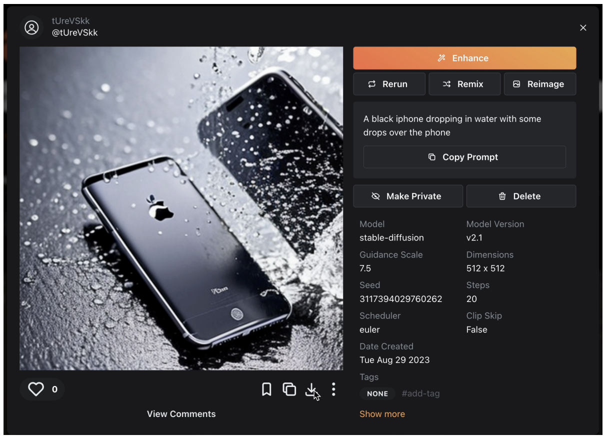
Original Video
I AM YOU ARE WE ARE AUSTRALIAN
Australians sporting culture brings Australia together as a nation!
Just out having fun
Playing a sport we love
Retrieval Result

Not Retrieved Correctly within Top-100.

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.

Retrieved Inorrectly 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.



A black iphone dropping in water with some drops over the phone

Model: stable-diffusion
Model Version: v2.1
Guidance Scale: 7.5
Dimensions: 512 x 512
Seed: 3117394029760262
Steps: 20
Scheduler: euler
Clip Skip: False
Date Created: Tue Aug 29 2023
Tags: NONE #add-tag

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

c. Silent
d. 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)?
   a. 0
   b. 1–10
   c. 10+
2. Have you ever used {brand} before?
   a. Yes
   b. No