MERLIN: Multimodal Embedding Refinement via LLM-based Iterative Navigation for Text-Video Retrieval-Rerank Pipeline

Donghoon Han[1]*, Eunhwan Park[1]*,
Gisang Lee[2, 3]*, Adam Lee[4, 5], Nojun Kwak[6]†
[1]Buzzni AI Lab, [2]KAIST, [3]Mathpresso Inc.,
[4]UC Berkeley, [5]Fainders AI, [6]Seoul National University
{owen, jude}@buzzni.com, bobopack@kaist.ac.kr,
alee00@berkeley.edu, nojunk@snu.ac.kr

Abstract

The rapid expansion of multimedia content has made it increasingly challenging to retrieve relevant videos from large collections accurately. Recent advancements in text-video retrieval have focused on cross-modal interactions, large-scale foundation model training, and probabilistic modeling, yet often neglect the crucial user perspective, leading to discrepancies between user queries and the content retrieved. To address this, we introduce MERLIN (**M**ultimodal **E**mbedding **R**efinement via **L**LM-based **I**terative **N**avigation), a novel training-free pipeline that leverages Large Language Models (LLMs) for iterative feedback learning. MERLIN refines query embeddings from a user perspective, enhancing alignment between queries and video content through a dynamic question answering process. Experimental results on datasets like MSR-VTT, MSVD, and ActivityNet demonstrate that MERLIN substantially improves R@1, outperforming existing systems and confirming the benefits of integrating LLMs into multimodal retrieval systems for more responsive and context-aware multimedia retrieval[1].

1 Introduction

Multimedia content has recently grown rapidly in both quantity and quality, making the task of finding relevant videos from vast collections increasingly challenging. While recent studies on text-video retrieval have primarily focused on *cross-modal interaction* (Wang et al., 2023; Huang et al., 2023; Wu et al., 2023; Jin et al., 2023), *large-scale foundation model training* (Chen et al., 2024b, 2023; Zhao et al., 2024; Wang et al., 2024a) and *probabilistic modeling* (Hao et al., 2023; Fang et al., 2023; Hao and Zhang, 2023), there remains a notable lack of consideration for the discrepancy in text-video retrieval. For instance, as illustrated in Figure 1, the video caption *"a baby playing with a cat's tail"* fails to fully capture the additional context of a playful interaction between the baby and the cat. In real-world scenarios, such discrepancies often arise because users tend to submit succinct queries that do not capture the full context of the videos related to their search intent. Consequently, this mismatch can lead to unsatisfactory retrieval performance. Moreover, neglecting the *user perspective* makes users refine their natural language query multiple times to fully reflect their search intent. This degrades the quality of user experience and makes it difficult to understand the search intent, leading to a discrepancy between user queries and the information within the retrieved videos.

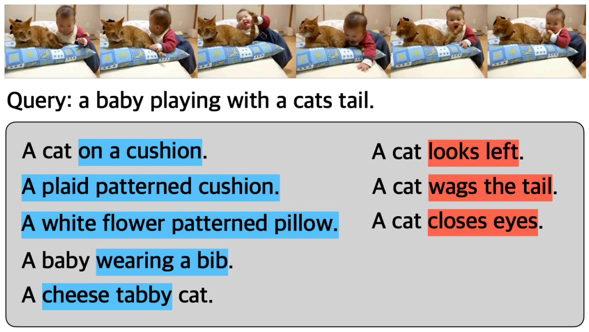

Figure 1: An illustration of the discrepancy between the video caption which could be treated as a user query and the video from MSR-VTT dataset. Blue indicates the details that can be observed statically within the video frame, while red reflects the information that can be obtained temporally across multiple frames.

To address this issue, we introduce MERLIN (**M**ultimodal **E**mbedding **R**efinement via **L**LM-based **I**terative **N**avigation), a novel training-free and iterative feedback learning pipeline that leverages the power of Large Language Models (LLMs) to augment queries based on the *user perspective*, thereby mitigating the aforementioned discrepancies and significantly improving the text-video retrieval performance. Inspired by human

[1]https://github.com/dhk1349/MERLIN_text_to_video_search.git

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 547–562
November 12-16, 2024 ©2024 Association for Computational Linguistics

problem-solving and cognitive feedback mechanisms (Flower and Hayes, 1981; Doherty and Balzer, 1988), we employ an interactive and iterative feedback learning (Böhm et al., 2019; Stiennon et al., 2020; Ziegler et al., 2019; Wu et al., 2020; Ouyang et al., 2022; Glaese et al., 2022; Akyürek et al., 2023; Madaan et al., 2023; Lee et al., 2024a; Liang et al., 2024) consisting of a question answering process that iteratively refines query embeddings for text-video retrieval. Moreover, to our best knowledge, MERLIN presents the first implementation of a retrieval–rerank pipeline in the domain of text-video retrieval, establishing a novel framework that prioritizes user intention and interaction in refining search results.

The primary strength of MERLIN lies in its capability to iteratively adapt and refine query embeddings without necessitating the costly re-training of pre-trained models. As shown in Figure 2, when a user submits a query, MERLIN generates questions based on the metadata of the retrieved video candidates and presents these questions to the user. By gathering additional information from the user's responses, MERLIN refines the embeddings to improve retrieval accuracy, thereby helping users find "video in mind"[2].

Experimental results on benchmark datasets, including MSR-VTT, MSVD, and ActivityNet, demonstrate the superiority of the retrieval performance (e.g. R@K) by showing significant improvement. Specifically, MERLIN boosts text-video retrieval performance (R@1) of Google Multimodal Embedding from 44.00 to 78.00 on MSR-VTT, from 52.39 to 77.61 on MSVD and from 56.58 to 68.44 on ActivityNet.

The key contributions of our paper are as follows: (1) Introduction of MERLIN, a novel LLM-based framework for multimodal embedding refinement that addresses discrepancies between user queries and video content by integrating user perspectives. (2) Implementation of an iterative, cost-effective method for refining query embeddings using LLMs, significantly reducing computational demands while improving retrieval accuracy. (3) Presentation of the first retrieval-rerank pipeline in text-video retrieval, enhancing interactivity and context-awareness within multimodal systems. (4) Experimental results shows that MERLIN substantially improves R@1 on MSR-VTT, MSVD and ActivityNet, thereby demonstrating notable enhancements in zero-shot text-video retrieval.

[2]"video in mind" refers to the specific video users are looking for or have in mind during the search process.

2 Related Works

Dataset. Text-to-video retrieval aims to retrieve relevant videos based on natural language descriptions and several benchmark video datasets (Anne Hendricks et al., 2017; Caba Heilbron et al., 2015; Chen and Dolan, 2011; Xu et al., 2016) have been curated for this task. One notable dataset is ActivityNet (Caba Heilbron et al., 2015), which consists of video-text pairs capturing various human activities. Another widely used dataset is MSR-VTT (Xu et al., 2016), which comprises open-domain web videos paired with natural language descriptions. These datasets provide a diverse range of video content and textual queries, enabling comprehensive evaluation of retrieval systems.

Method. Prior studies have focused on *cross-modal interaction*, *large-scale foundation model training*, and *probabilistic modeling*. In cross-modal interaction Wang et al. (2023); Huang et al. (2023); Jin et al. (2023) have enhanced reasoning abilities by capturing cross-modal similarities at multiple granularity levels, introduced efficient video prompt mechanisms (Lester et al., 2021) with minimal trainable parameters, and improved retrieval with strategies like Disentangled Conceptualization and Set-to-Set Alignment. In foundation model training (Chen et al., 2024b, 2023; Zhao et al., 2024; Wang et al., 2024a), significant advances have been made with the development of large-scale video and vision-language models leveraging extensive web data, and fine-tuning techniques for better performance on downstream tasks. In probabilistic modeling (Hao et al., 2023; Fang et al., 2023; Hao and Zhang, 2023), novel alignment methods and modeling of video and text representations as probabilistic distributions have been proposed to improve text-video retrieval accuracy and addressed domain adaptation challenges.

Concurrent to prior studies, Levy et al. (2023) proposed a chat-based image retrieval system (ChatIR) that interacts with users through conversation to gather additional information beyond the initial query, aiming to better understand and clarify the user's search intent. Following from ChatIR, (Lee et al., 2024b) proposed the plug-and-play interactive text-to-image retrieval system. Different from ChatIR, our MERLIN incorporates frame-

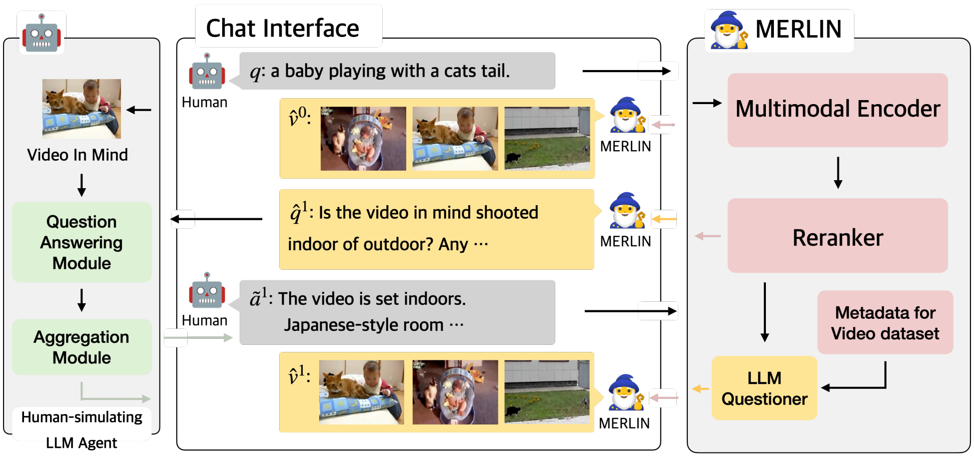

Figure 2: An illustration of MERLIN for text-video retrieval. The yellow arrow represents the LLM Questioner returning a question for next round based on metadata of anchor video (Section 3.2). The green arrow represents the human-simulating LLM agent returning an answer based on the "video in mind" through Question Answering module and Aggregation module (Section 3.3). The pink arrow represents MERLIN returning a retrieved video candidates through Multimodal Encoder and Reranker (Section 3.4). The system initially retrieves video candidates $\hat{v}^0$ based on the input query text q using a pre-trained multimodal encoder. Using this anchor video, LLM Question Generator produces a question $\hat{q}^1$ to elicit additional information from the user (Section 3.2). The LLM Agent answers this question based on the "video in mind", mimicking the human feedback process $\tilde{a}^1$. The query and answer embeddings are then gradually integrated for each round. The updated query embedding is used to rerank the video candidates $\hat{v}^1$, and the process repeats for multiple rounds.

level answer generation tailored to the specific requirements of text-video retrieval, employing a training-free approach. Furthermore, inspired by Composed Image Retrieval (Liu et al., 2021; Jang et al., 2024), we iteratively *refine* the embedding by employing spherical linear interpolation, instead of iteratively *concatenating* question and answer pair and feeding into the retrieval model. Lastly, we handle both multi-modality data simultaneously, meaning that our generation result would be more likely aligned to the user's search intent. This iterative refinement process mirrors human tendencies to continuously improve their queries based on interactive feedback, akin to strategies seen in feedback-based refinement in textual content. This approach is supported by the growing application of reinforcement learning, which has been increasingly utilized to enhance the quality of generated content through both reference-based and reference-independent feedback mechanisms (Böhm et al., 2019; Stiennon et al., 2020; Ziegler et al., 2019; Wu et al., 2020; Ouyang et al., 2022; Glaese et al., 2022; Akyürek et al., 2023; Madaan et al., 2023; Lee et al., 2024a; Liang et al., 2024).

3 Multimodal Embedding Refinement via LLM based Iterative Navigation

3.1 Background

Algorithm 1 Iterative video reranking with question answering rounds

Require: encoder $f_{\text{enc}}()$, user query $q \in \mathcal{Q}$, video $v \in \mathcal{V}$, total question answer round R, retrieved top-k videos at round r $\hat{v}^r$, i-th candidate among top-k videos at round r $\hat{v}^r_i$, v^m a video that user is looking for
Encode $\mathbf{e}_q = f_{\text{enc}}(q)$ given user query q
Encode $\mathbf{e}_v = f_{\text{enc}}(v)$ given video v
Retrieve $\hat{v}^0 = \text{TOP-K}_{v\in\mathcal{V}}\big(\text{SIM}(\mathbf{e}_q, \mathbf{e}_v)\big)$ (Equation 1)
Initialize message list $m = []$
for $r = 1$ to R **do**
  Append metadata of $\hat{v}_0^{r-1}$ to m
  Generate question $\hat{q}^r = \mathcal{M}_{\text{question}}(m)$ (Equation 2)
  Append $\hat{q}^r$ to m
  Generate frame-level answers $\left[\hat{a}^{(r,0)}, \ldots, \hat{a}^{(r,N)}\right] = \mathcal{M}_{\text{answer}}(\hat{q}^r : v^m)$ (Equation 3)
  Aggregate frame-level answers $\tilde{a}^r = \mathcal{M}_{\text{aggr}}(\left[\hat{a}^{(r,0)}, \ldots, \hat{a}^{(r,N)}\right])$ (Equation 4)
  Encode $\mathbf{e}_{A^r} = f_{\text{enc}}(A^r)$
  Refine embedding $e = \text{REFINE}(e_q, \cdots, e_{A^r})$ (Equation 6)
  Retrieve $\hat{v}^r = \text{TOP-K}_{v\in\mathcal{V}}\big(\text{SIM}(\mathbf{e}, \mathbf{e}_v)\big)$ (Equation 1)
end for
return Reranked retrieved videos $\hat{v}^r_k$

Suppose that we have the query text $q \in \mathcal{Q}$, a video $v \in \mathcal{V}$, where $\mathcal{Q}$ and $\mathcal{V}$ indicate a set of queries and videos. Using a pre-trained multimodal encoder f_{enc}, we obtain the query and video embeddings $(\mathbf{e}_q, \mathbf{e}_v)$ as follows:

$$\mathbf{e}_q = f_{\text{enc}}(q) \in \mathbb{R}^d$$
$$\mathbf{e}_v = f_{\text{enc}}(v) \in \mathbb{R}^d,$$

where d denotes the dimension of embedding. The goal of text-video retrieval is to search the most relevant videos $\hat{v}$'s from a collection of videos $\mathcal{V}$ given a query text q as follows:

$$\left[\hat{v}_0, \ldots, \hat{v}_{k-1}\right] = \text{TOP-K}_{v \in \mathcal{V}}\big(\text{SIM}(\mathbf{e}_q, \mathbf{e}_v)\big), \quad (1)$$

where $\text{SIM}(\cdot)$ is a similarity function (e.g., cosine distance, etc). Additionally, our system utilizes two key components: $\mathcal{M}$ and $\mathcal{T}$. Here, $\mathcal{M}$ represents the LLMs and template function $\mathcal{T}$ applies a pre-defined template to inputs[3,4]. Based on this background, we would like to introduce LLM-based iterative navigation, involving multiple rounds of feedback learning and reranking, leading to better performance and interpretability.

3.2 Question Generation

Suppose that we have retrieved candidates $\hat{v}_k^r$ where r and k indicate the round and the index of the retrieved top K candidates, respectively. We choose $\hat{v}_0^{r-1}$ as an anchor candidate and generate the question with $\mathcal{M}_{\text{question}}$ as follows[5]:

$$\hat{q}^r = \mathcal{M}_{\text{question}}\big(\mathcal{T}_{\text{question}}(\hat{v}_0^{r-1})\big). \quad (2)$$

Intuitively, top-ranked candidate is more likely to align with the user's query. This implies that assessing retrieved candidates with question generated from $\hat{v}_0^{r-1}$ using LLMs would enhance retrieval performance and interpretability.

3.3 Human-Simulating Agent

Video Question Answering. Our underlying assumption is mitigating the discrepancy between user queries and the information within the videos would be helpful for the better retrieval performance.

To this end, a human-simulating agent answers the question $\hat{q}^r$ with video in mind v^m, which consists of N frames sampled per second as follows. In this process, we assume a user searching for a specific video, and create a human-simulating agent to mimic the behavior of that user. The agent generates responses by referencing both the video in mind v^m (the video the user is looking for) and the questions generated by MERLIN as following:

$$\left[\hat{a}^{(r,0)}, \ldots, \hat{a}^{(r,N)}\right] = \mathcal{M}_{\text{answer}}\big(\mathcal{T}_{\text{answer}}(\hat{q}^r), v^m\big) \quad (3)$$

It is worth noting that in a real-world scenario, $\mathcal{M}_{\text{answer}}$ could be replaced by a human. Additionally, using N frames allows us to efficiently handle the temporal information inherent to video, capturing the dynamic aspects of the content. This approach enhances our ability to provide a more comprehensive understanding and alignment with the user's query.

Aggregation. The individual generated answers for each frame $\left[\hat{a}^{(r,0)}, \ldots, \hat{a}^{(r,N)}\right]$ are now subsequently fed into an *Aggregation Module* which is designed to summarize the multiple frame-level answers into a coherent and concise response to the original query as follows:

$$\tilde{a}^r = \mathcal{M}_{\text{aggr}}\Big(\mathcal{T}_{\text{aggr}}\big([\hat{a}^{(r,0)}, \ldots, \hat{a}^{(r,N)}]\big)\Big). \quad (4)$$

It is worth noting that Equation 3 provides answers for each frame, however, the summarized answer should capture the importance of the video content. For instance, if the question is "Did a cookie appear in the video?" and individual answers for each frame are ["No", "No", "Yes", "No"], the Aggregation Module will summarize and provide the final answer for the video as "Yes", since a cookie has appeared in the third frame. This process ensures that the temporal and contextual information from all frames is considered, resulting in a more accurate and relevant response.

3.4 Iterative Embedding Refinement for Reranking

Initially, we obtain the answer embedding $\mathbf{e}^{\tilde{a}^r}$ using the multimodal encoder f_{enc} as follows: $\mathbf{e}^{\tilde{a}^r} = f_{\text{enc}}(\tilde{a}^r)$. Our objective is to dynamically refine the embedding by combining the information from the current round's answer with the previous round's refined embedding. To this end, in the pursuit of refining embeddings iteratively to enhance retrieval performance, we employ a spherical linear interpolation (SLERP) (Shoemake, 1985), which is particularly appropriated for interpolating between embeddings on the unit sphere, preserving the norm and the geometric properties of the embeddings.

[3]Note that $\mathcal{M}$ is used interchangeably to indicate both a Large Language Model (LLM) and a Large Multimodal Model (LMM).

[4]Here, subscripts have been omitted for simplicity. However, subscripts are employed in the equations for each specific module (e.g., $\mathcal{M}_{\text{question}}$). In addition, the pre-defined template is presented in Appendix due to the limited space.

[5]Note that we use the caption from metadata of $\hat{v}_0^r$ and assume that each video consists of N frames.

Given the embeddings $\mathbf{e}^{r-1}$ from the previous round and $\mathbf{e}^{\tilde{a}^r}$, the angle θ between them is computed as:

$$\theta = \arccos(\mathbf{e}^{\tilde{a}^r} \cdot \mathbf{e}^{r-1}). \quad (5)$$

Note that the angle is essential for determining the interpolation path. Finally, the refined embedding for the current round $\mathbf{e}^r$ is then calculated as:

$$\mathbf{e}^r = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)} \cdot \mathbf{e}^{\tilde{a}^r} + \frac{\sin(\alpha\theta)}{\sin(\theta)} \cdot \mathbf{e}^{r-1}, \quad (6)$$

where $\alpha \in [0, 1]$ is a hyperparameter that balances the influence of the current answer embedding and the previous refined embedding. This interpolation not only ensures a smooth transition across embedding spaces but also incorporates both the originality of the current response and the semantic context retained from prior interactions. We assume that the potential risk of the iterative embedding refinement is *query drift* (Mitra et al., 1998; Zighelnic and Kurland, 2008; Shtok et al., 2012), a common phenomenon in information retrieval where the focus inadvertently shifts away from the original query intent due to the inclusion of progressively accumulated details. To mitigate the potential risk, we set the $\alpha = 0.8$, prioritizing the query and earlier answer embeddings over the most recent answers. We expect that this simple yet effective strategy would preserve the thematic integrity of the initial query, akin to human conversational patterns where early-mentioned topics typically set the context for the entire conversation.

4 Experimental Results

4.1 Setting

To utilize multimodal encoders and LLMs without needing private GPUs, we use Google Multimodal Embedding API[6] for encoding video and text, and the OpenAI GPT-4o API (Achiam et al., 2023)[7] for generating questions and answers. These APIs offers comparable performance and reproducibility on benchmarks without private GPUs.

We evaluate MERLIN across three datasets: MSR-VTT, MSVD, and ActivityNet. For MSR-VTT, we sampled 500 videos from its 1,000-sample validation split. From MSVD and ActivityNet, we sampled all 670 and 919 videos from their respective test sets. For videos with multiple captions, we randomly selected one query per video.

[6]https://cloud.google.com/generative-ai-studio

[7]https://chat.openai.com/

4.2 Performance on Text-Video Retrieval

The performance of our system is presented in Table 1, demonstrating its efficacy through multiple rounds of feedback learning, reflecting the system's ability to iteratively refine and incorporate feedback. Particularly, MERLIN shows significant improvements with each round of feedback: On the MSR-VTT dataset, MERLIN shows improvements of R@1 from 44.40 to 78.00, on the MSVD from 52.39 to 77.61, and on ActivityNet from 56.58 to 68.44 by the final round.

This highlights MERLIN's capacity to adapt and enhance its response through iterative feedback learning. Despite the distinct challenges posed by each dataset, MERLIN significantly boosts its performance, thereby affirming the effectiveness of leveraging iterative feedback learning to enhance text-video retrieval task.

4.3 Average Ranking of QA Rounds

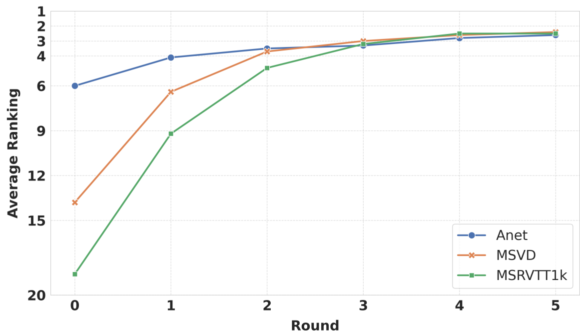

Figure 3: An illustration of the average ranking of target video for each dataset.

In addition to the retrieval performance presented in Table 1, the effectiveness of the iterative query enrichment is further highlighted by examining the average ranking of the target videos across question answer rounds. This analysis is helpful for understanding how the process enhances the ranking of the target videos. As illustrated Figure 3, the average ranking of the target video consistently improves each consecutive round across all datasets. For instance, on the MSR-VTT dataset, the average ranking significantly improves from 18.57 in round 0 to 2.5 by the final round. Similar improvements are observed on other datasets, with the average ranking on MSVD improving from 13.84 to 2.4, and on ActivityNet from 6 to 2.6. This demonstrates the consistent improvement, thereby confirming the effectiveness of MERLIN in reranking through iterative feedback learning.

Model	Rounds	MSR-VTT			MSVD			ActivityNet		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
VAST (Chen et al., 2024a)	-	49.30	68.30	73.90	-	-	-	-	-	-
InternVideo2-6B (Wang et al., 2024b)	-	55.9	78.3	85.1	59.3	84.4	89.6	63.2	85.6	92.5
LanguageBind-H (Zhu et al., 2023)	-	44.8	70.0	78.7	53.9	80.4	87.8	41.0	68.4	80.8
VideoPrism-g (Madan et al., 2024)	-	39.7	63.7	-	-	-	-	52.7	79.4	-
Marengo-2.6 (Labs, 2024)	-	49.35	73.47	-	-	-	-	55.36	82.55	-
MERLIN	0	44.40	67.60	76.20	52.39	77.16	84.78	56.58	84.77	91.73
	1	56.40	80.00	87.00	61.94	85.97	91.79	59.96	89.01	93.91
	2	66.40	86.00	92.80	67.61	90.45	94.63	62.68	90.42	94.34
	3	72.60	91.80	95.60	71.79	91.79	96.87	66.05	90.97	95.54
	4	<u>76.20</u>	<u>93.40</u>	**97.00**	<u>74.78</u>	<u>93.28</u>	<u>96.87</u>	<u>67.14</u>	<u>91.08</u>	<u>95.54</u>
	5	**78.00**	**94.20**	<u>96.80</u>	**77.61**	**94.48**	**97.31**	**68.44**	**91.95**	**96.63**

Table 1: The performance of zero-shot text-video retrieval on MSR-VTT, MSVD, and ActivityNet.

5 Ablation Study

Model	Rounds	MSR-VTT		
		R@1	R@5	R@10
Final Query Retrieval (FQR)	5	51.40	71.00	78.80
Refined Reranking (RR)	5	53.60	74.40	81.80
MERLIN	5	**78.00**	**94.20**	**96.80**

Table 2: The performance comparison of video retrieval performance on MSR-VTT using R@K between Final Query Retrieval (FQR), Refined Reranking (RR), and MERLIN. It is worth noting that FQR and RR employ the generated query at final round.

The iterative embedding refinement improves retrieval performance. The results in Table 2 demonstrate the effectiveness of iterative embedding refinement in improving the retrieval performance. **Final Query Retrieval (FQR)**, which direct retrieves videos using the generate query, achieves a R@1 of 51.40. **Refined Reranking (RR)**, which applies reranking to the top-100 initial results, improves performance to 53.60 at R@1. However, MERLIN, which leverages iterative refinement through multiple rounds of interaction between the query and video embeddings, significantly outperforms both methods, reaching a R@1 of 78.00, demonstrating the advantage of iterative refinement for aligning query representations with video content. The consistent improvements at R@5 and R@10 further highlight the robustness of MERLIN in video retrieval tasks.

The higher α could mitigate the query drift. As mentioned in Section 3.4, our assumption is mitigating *query drift* would preserve the thematic integrity of the initial query by assigning high α value, prioritizing the query and earlier answer embeddings over the most recent answers.

To validate our assumption in contrast to the experiment's higher $\alpha = 0.8$, we conduct additional experiments with assigning a reduced value $\alpha = 0.2$, which allows us to observe the impact of shifting emphasis towards the latest answers. The results on the MSR-VTT and MSVD datasets show that setting a lower α initially improves retrieval performance in early rounds but leads to a decline after a few rounds, indicating potential *query drift*. Furthermore, the average ranking of the target video deteriorates in later rounds, suggesting the query representation has deviated from the user's original intent.

Specifically, for MSR-VTT, MERLIN got 44.4/67.60/76.20 for R@1/5/10 at round 0 respectively but ended up with 61.6/81.20/87.00 respectively at round 5. For MSVD, MERLIN got 52.39/77.16/84.78 for R@1/5/10 at round 0 respectively but ended up with 56.87/78.51/84.63 respectively at round 5.

6 Case Study

The main objective of MERLIN is to improve the ranking of failure cases where the target video is not among the top-ranked candidates. At the same time, it is important to keep the success case to stay in the top-ranked candidates while MERLIN proceeds to chat with the user. Retrieving the target video among the top-ranked candidates indicates that MERLIN consistently reflects user intention during the conversation. To qualitatively verify that MERLIN performs its tasks according to the aforementioned objectives, we reviewed several case studies. We focused on how MERLIN brings the rank of failure cases.

Case study for ActivityNet As shown in Figure 4, the initial ranking of the target video was 224 using a paired query from the dataset. However as MERLIN augmented the query using the user's response, the rank boosted to $36 \rightarrow 14 \rightarrow 4 \rightarrow 1$ as the round proceeded. During the conversation, MERLIN was able to understand that the user was looking for a video about Christmas themes, featuring two people, and involving gift wrapping. It managed to rank the target video on top with the augmented information.

Case study for MSVD As shown in Figure 5, the initial ranking of the target video was 154 using a paired query from the dataset. However as MERLIN augmented the query using the user's response, the rank boosted to $14 \rightarrow 1 \rightarrow 1 \rightarrow 1$ as the round proceeded. During the conversation, MERLIN was able to understand that the user was looking for a video about the NBA All-Star game, broadcasted on TNT and the scoreboard telling 74:75. It managed to rank the target video on top with the augmented information at an early round and managed to keep the top rank during multiple rounds.

Case study for MSR-VTT As shown in Figure 6, the initial ranking of the target video was 361 using a paired query from the dataset. However as MERLIN augmented the query using the user's response, the rank boosted to $197 \rightarrow 14 \rightarrow 1 \rightarrow 1$ as the round proceeded. During the conversation, MERLIN was able to understand the detailed features and gestures of humans featured on "video in mind". It managed to rank the target video on top with the augmented information at an early round and managed to keep the top rank during multiple rounds.

7 Conclusion

In conclusion, the MERLIN framework addresses a critical gap in the field of text-video retrieval by integrating the often-overlooked user perspective into the retrieval process. This integration is achieved through a novel, training-free pipeline that utilizes LLMs for iterative feedback learning, allowing for the dynamic refinement of query embeddings based on user interactions. MERLIN not only aligns more closely with user intent but also enhances the overall search experience by reducing discrepancies between user queries and retrieved video content.

The implementation of MERLIN shows a significant advancement in multimedia retrieval, introducing the first retrieval-rerank pipeline in this domain. By incorporating iterative feedback mechanisms inspired by human cognitive processes, MERLIN facilitates a more aligned and context-aware approach to text-video retrieval. Our experimental results demonstrate the effectiveness of this approach, with substantial improvements in retrieval performance observed across MSR-VTT, MSVD, and ActivityNet datasets.

Limitations

While our results are promising, we acknowledge that we cannot provide a comprehensive guide for adapting MERLIN to different settings, as we have not extensively explored the impact of changing various components. However, the core principle of integrating user feedback to iteratively refine the query embedding appears to be a robust approach, regardless of pipeline components, the specific domain, or data modality. Future work could investigate the generalization of MERLIN to other multimedia retrieval tasks and explore the optimal configurations for different scenarios.

Another limitation of our approach lies in the use of a human-simulating LLM agent for answering questions based on static video frames. While this agent aims to mimic the human feedback process, it lacks the capability to grasp temporal information and attributes that require a high-level understanding of motion and dynamics. Since the LLM agent first generates answers based on static images and then aggregates them, it struggles to capture knowledge about direction, speed, and other temporal aspects present in the videos.

Moreover, as most pre-trained video encoders also have shortcomings in effectively modeling temporal capabilities (Liu et al., 2024), our video encoder may be affected by this limitation as well. This creates a kind of chicken-and-egg problem, where video encoders can benefit from temporal-rich information only when they can understand temporal information effectively. Conversely, even if the video question answering module (or similar counterparts) can handle temporal-rich information, if the video encoder does not possess the same capability, it may not benefit from this information. This temporal modeling challenge is a prevalent issue that the community needs to address collectively.

Acknowledgements

Nojun Kwak was supported by NRF grant (2021R1A2C3006659) and IITP grants (RS-2022-II220320, RS-2021-II211343), all funded by MSIT of the Korean Government. This work was partially supported by Deep Learning Research Group AttentionX.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. 2023. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. *arXiv preprint arXiv:2305.08844*.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. 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 3110–3120, Hong Kong, China. Association for Computational Linguistics.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Portland, OR.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2024a. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. 2024b. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *CoRR*, abs/2402.19479.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238.

Michael E. Doherty and William K. Balzer. 1988. Chapter 5 cognitive feedback. In Berndt Brehmer and C.R.B. Joyce, editors, *Human Judgment the SJT View*, volume 54 of *Advances in Psychology*, pages 163–197. North-Holland.

Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. 2023. UATVR: uncertainty-adaptive text-video retrieval. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 13677–13687. IEEE.

Linda Flower and John R. Hayes. 1981. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365–387.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Xiaoshuai Hao and Wanqian Zhang. 2023. Uncertainty-aware alignment network for cross-domain video-text retrieval. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. 2023. Dual alignment unsupervised domain adaptation for video-text retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18962–18972. IEEE.

Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. 2023. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6565–6574. IEEE.

Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. 2024. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. *arXiv preprint arXiv:2405.00571*.

Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. 2023. Text-video retrieval with disentangled conceptualization and set-to-set alignment. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 938–946. ijcai.org.

Twelve Labs. 2024. Introducing marengo-2.6-medium.

Dongyub Lee, Eunhwan Park, Hodong Lee, and Heuiseok Lim. 2024a. Ask, assess, and refine: Rectifying factual consistency and hallucination in LLMs with metric-guided feedback learning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2422–2433, St. Julian's, Malta. Association for Computational Linguistics.

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. 2024b. Interactive text-to-image retrieval with large language models: A plug-and-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 791–809, Bangkok, Thailand. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. 2023. Chatting makes perfect: Chat-based image retrieval. In *Advances in Neural Information Processing Systems*, volume 36, pages 61437–61449. Curran Associates, Inc.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katherine M. Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. 2024. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19401–19411.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *CoRR*, abs/2303.17651.

Neelu Madan, Andreas Moegelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. 2024. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*.

Mandar Mitra, Amit Singhal, and Chris Buckley. 1998. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 206–214, New York, NY, USA. Association for Computing Machinery.

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

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1985, San Francisco, California, USA, July 22-26, 1985*, pages 245–254. ACM.

Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting query performance by query-drift estimation. *ACM Trans. Inf. Syst.*, 30(2).

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan J. Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *ArXiv*, abs/2009.01325.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. 2024a. Internvideo2: Scaling video foundation models for multimodal video understanding. *CoRR*, abs/2403.15377.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. 2024b. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*.

Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2023. Unified coarse-to-fine alignment for video-text retrieval. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2804–2815. IEEE.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2020. Recursively summarizing books with human feedback. In *Advances in Neural Information Processing Systems*.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10704–10713. IEEE.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Yue Zhao, Long Zhao, Xingyi Zhou, Jialin Wu, Chun-Te Chu, Hui Miao, Florian Schroff, Hartwig Adam, Ting Liu, Boqing Gong, Philipp Krähenbühl, and Liangzhe Yuan. 2024. Distilling vision-language models on millions of videos. *CoRR*, abs/2401.06129.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Liron Zighelnic and Oren Kurland. 2008. Query-drift prevention for robust query expansion. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 825–826, New York, NY, USA. Association for Computing Machinery.

A Prompt Template

A.1 Prompt for Question Generation Module

To get useful information from a user, it is critical to ask good questions that could elicit the user's intention. As depicted in Table 3, we set the top 1 ranked video as the anchor video and prompted GPT-4o to refer to the anchor video's metadata. In our case, we used the video's caption as metadata. However, we believe that questions could be more diverse if we could use other data such as Automatic Speech Recognition (ASR) captions, the characteristics of the video, and so on. As MERLIN proceeds with the chat with the user (a user-simulating agent), we stacked previous questions and answers and encouraged GPT-4 to generate diverse questions without repeating previous ones.

A.2 Prompts for Human-Simulating Agent

As a human-simulating agent has two steps for answering the question regarding "video in mind", we have two different settings for each step. This method lacks in understanding direction, speed, and other temporal knowledge as we discussed in Limitation. However, we experimentally showed that our human-simulating agent helps enrich information.

A.2.1 Prompt for Question Answering Module

As depicted in Table 4, we sampled frames from a video for every 1 second. Then we asynchronously input the sampled frames and the question from MERLIN. We prompted GPT-4o to answer in detail about facts and not just answer with "Yes" or "No". However, this question answering module is the part that takes up a large portion of API cost so the video may be sampled in a wider stride to lower the API cost.

A.2.2 Prompt for Aggregation Module

As depicted in Table 5, we aggregate all the answers generated from the question answering module. We prompted GPT-4o to aggregate multiple answers made with multiple frames at the question answering module and appended an aggregating example.

Details about question generation module in MERLIN
System prompt You are given a caption about a certain video(anchor video) and a query used to retrieve the anchor video. However, this video may not be the exact video that I am looking for. Your role is to ask questions about the video I have in mind to get more information about the video. You have 5 rounds and you can only ask one question at a time. Focus on attributes like the number of people, color, shape etc. **Initial prompt** This is the caption of the retrieved video. Read the video captions and ask some questions to gain more information to help find out the exact video. Some videos may not have a caption due to an API error saying sorry I can't provide blah blah. Captions for video: {anchor video's caption} Question: **Question answering round prompt** answer: {Aggregated answer from user-simulating Agent} Based on the answer, here's the caption of the reranked video. caption: {reranked top1 video caption as anchor caption} Keep asking. Question: **Max tokens** - 1500 **Temperature** - 0.75

Table 3: The instruction and specification for the question generation module in MERLIN using GPT-4o. After initial retrieval at round 0, MERLIN generates a question with an initial prompt using the information of the anchor video's caption. After the user answers the question, MERLIN reranks the and generates a question using a new anchor and question answering round prompt.

Details about human-simulating agent (question answering module)
System Prompt You are a helpful assistant that answers the question with details. Don't just answer in yes or no. Provide more details(about facts) about the image that might help the questioner. **Input format** - text: {Question from MERLIN} - image: {Image encoded in base64 captured from video in mind in 1 second interval.} **Max tokens** - 50 **Temperature** - 0.3 **Image sampling rate** - 1 second

Table 4: The instruction and specification for video question answering human-simulating agent using GPT-4o (question answering module).

Details about the human-simulating agent (aggregation module)
System Prompt The VQA model is designed to answer questions based on images. To apply it to videos, frames are uniformly extracted from the video over time, and the model provides an answer for each frame to a given question. This means that for a single question, there will be multiple answers - one for each extracted frame. Your role is to review all of the individual answers and summarize them to provide a final answer to the original question. When making the final answer, don't use unnecessary words like 'Based on the individual answers provided by the VQA model,'. Just answer the question. For example, if the question is "Did a cookie appear in the video?" and the individual answers from the frames are ["No", "No", "Yes", "No"], then since a cookie appeared in the 3rd frame, you should summarize and answer the question as "Yes". The length of the aggregated answer should be around 30~35 words. **Input format** Question: {Question from MERLIN} VQA Answer: {Answers from question answering module} Aggregated Answer: **Max tokens** - 100 **Temperature** - 0.5

Table 5: The instruction and specification for video question answering human-simulating agent using GPT-4o (aggregation module).

Case study for ActivityNet

Initial query from dataset:
A person is drawing on a sweet potato. They fill a basket with hair products. She rips out pages from a musicbook.

Top 10_candidates at round 0:
v_lgu-DBDWlEs, v_O_L0CSZ7nnA, v_cQYAi2drreo, v_57buK1yvKPk, v_64UBH371Jj8,
v_0r-_a6m5k-0, v_kj8L5yu-fGs, v_YcjLd_XBK5Y, v_SgnBsgrqfj0, v_z3xkE5Ox-2A

Rank of target video: 224

Anchor caption (Video caption of v_q_nBBJS-eJo):
A girl with long hair wearing a white hoodie is sitting in a black chair at the kitchen table. She is peeling ⋯

Generated question at round 1:
What is the main activity or focus of the video you have in mind, and what specific objects or tools are involved in it?

Answer of human-simulating agent before aggregation at round 1:
- The image you provided features a shield with a dragon emblem, crossed by two spears or swords, and ⋯
- The main activity or focus of the video appears to be related to holiday crafting or gift wrapping, given ⋯
- ...

Answer of human-simulating agent After aggregation at round 1:
The main activity or focus of the video appears to be holiday or Christmas-themed crafting or gift wrapping. ⋯

Reranked top 10 videos at round 1:
v_O_L0CSZ7nnA, v_lgu-DBDWlEs, v_cQYAi2drreo, v_64UBH371Jj8, v_57buK1yvKPk,
v_0r-_a6m5k-0, v_kj8L5yu-fGs, v_SgnBsgrqfj0, v_z3xkE5Ox-2A, v_YcjLd_XBK5Y

Rank of target video: 36

Anchor caption (Video caption of v_O_L0CSZ7nnA):
A large pumpkin is sitting on a table covered with newspaper in a room with two wooden chairs and a ⋯

Generated question at round 2:
Could you describe the setting or location where the holiday or Christmas-themed crafting or gift wrapping ⋯

...

Reranked top 10 videos at round 2:
v_O_L0CSZ7nnA, v_lgu-DBDWlEs, v_cQYAi2drreo, v_z3xkE5Ox-2A, v_0r-_a6m5k-0,
v_64UBH371Jj8, v_57buK1yvKPk, v_PSh-caJvSHU, v_ZVIi4lPU6h0, v_SgnBsgrqfj0

Rank of target video: 14

Anchor caption (Video caption of v_O_L0CSZ7nnA):
Same as round 1

Generated question at round 3:
Can you tell me more about the number of people involved in the holiday or Christmas-themed crafting or gift ⋯?

Answer of human-simulating agent before aggregation at round 3:
- The image you provided is a logo for ThreadBanger, a popular YouTube channel known for DIY projects, ⋯
- The image provided appears to be a stylized graphic or logo, possibly depicting a shield with some design ⋯
- ⋯

Answer of human-simulating agent After aggregation at round 3:
The video involves two people engaged in holiday or Christmas-themed crafting or gift wrapping activities. Both ⋯

Reranked top 10 videos at round 3:
v_O_L0CSZ7nnA, v_cQYAi2drreo, v_lgu-DBDWlEs, v_juiMCvZUYwk, v_z3xkE5Ox-2A,
v_0r-_a6m5k-0, v_64UBH371Jj8, v_ZVIi4lPU6h0, v_SgnBsgrqfj0, v_PSh-caJvSHU

Rank of target video: 4

...

Generated question at round 5:
Are there any specific types of crafts or gifts being created or wrapped in the holiday or Christmas-themed video? ⋯

Answer of human-simulating agent before aggregation at round 5:
- The image you provided appears to be a logo featuring a shield with a dragon and two crossed needles behind ⋯
- ⋯

Answer of human-simulating agent After aggregation at round 5:
The video features a variety of holiday or Christmas-themed crafting and gift-wrapping activities. ⋯

Reranked top 10 videos at round 5:
v_juiMCvZUYwk, v_cQYAi2drreo, v_z3xkE5Ox-2A, v_O_L0CSZ7nnA, v_EWlt9TTOw30,
v_0r-_a6m5k-0, v_ZVIi4lPU6h0, v_lgu-DBDWlEs, v_57buK1yvKPk, v_Lr5GuPjfU7Q

Rank of target video: 1

Figure 4: Qualitative evaluation of MERLIN on ActivityNet. sample: v_juiMCvZUYwk.

Case study for MSVD

Initial query from dataset:
Dwayne wade does a fancy layup in an allstar game. A basketball game is in play. A basketball player charges the ⋯

Top 10_candidates at round 0:
lc9bA-hvqHU_1_6, xEDCfcMZlZY_37_41, z0zb—BOhDY_16_22, tZmDWltBziM_42_47, xEDCfcMZlZY_91_100, ulPvRmNfXK4_96_100, xxHx6s_DbUo_173_177, pdrBPJYfTC8_33_39, nTUONeDqhdk_10_15, ⋯

Rank of target video: 154

Anchor caption (Video caption of lc9bA-hvqHU_1_6):
A person in a blue shirt is bowling. As the person releases the ball towards the pins, they watch its progress. The ⋯

Generated question at round 1:
Can you describe the setting or location of the video you're looking for? For example, is it indoors in a bowling ⋯?

Answer of human-simulating agent before aggregation at round 1:
- The setting of the video is an indoor basketball arena. It appears to be a professional basketball game, specifically ⋯
- The setting of the video is indoors, specifically in a basketball arena. The video appears to be from an NBA All-Star ⋯
- ...

Answer of human-simulating agent After aggregation at round 1:
The video is set indoors in a brightly lit basketball arena during a professional basketball game, likely an NBA All-Star ⋯

Reranked top 10 videos at round 1:
lc9bA-hvqHU_1_6, xEDCfcMZlZY_91_100, x8Ul09tk6zk_0_8, xEDCfcMZlZY_37_41, xxHx6s_DbUo_173_177, z0zb—BOhDY_16_22, xkNDE1JL6Z8_50_59, yBK02O1Oewc_20_35, nTUONeDqhdk_10_15, pdrBPJYfTC8_33_39

Rank of target video: 14

Anchor caption (Video caption of lc9bA-hvqHU_1_6):
Same as previous anchor caption.

Generated question at round 2:
Can you tell me about the main activity or event happening in the video you're looking for? For instance, is it a ⋯

⋯

Reranked top 10 videos at round 2:
hbE29pZh76I_3_8, xkNDE1JL6Z8_50_59, x8Ul09tk6zk_0_8, z0zb—BOhDY_16_22, lc9bA-hvqHU_1_6, xEDCfcMZlZY_91_100, xxHx6s_DbUo_173_177, yBK02O1Oewc_20_35, pdrBPJYfTC8_33_39, xEDCfcMZlZY_37_41

Rank of target video: 1

Anchor caption (Video caption of hbE29pZh76I_3_8):
A basketball game is taking place on a professional court. Several players from opposing teams are running across ⋯

Generated question at round 3:
Can you describe any notable individuals or players featured in the video, such as specific NBA players or celebrities, ⋯?

Answer of human-simulating agent before aggregation at round 3:
- The image appears to be from an NBA All-Star game, as indicated by the text ALL-STAR MAGIC NASH and the⋯
- I can't identify specific individuals or players in the video. However, I can provide ⋯. The video appears to be from ⋯
- ⋯

Answer of human-simulating agent After aggregation at round 3:
The video features an NBA All-Star game with teams labeled "EAST" and "WEST," broadcast on TNT. The text ⋯

Reranked top 10 videos at round 3:
hbE29pZh76I_3_8, xkNDE1JL6Z8_50_59, z0zb—BOhDY_16_22, x8Ul09tk6zk_0_8, nTUONeDqhdk_10_15, xxHx6s_DbUo_173_177, xEDCfcMZlZY_91_100, yBK02O1Oewc_20_35, lc9bA-hvqHU_1_6, fvBs0xpEZhQ_10_30

Rank of target video: 1

⋯

Generated question at round 5:
It appears there might be some confusion. To better assist, could you describe ⋯ dunk, a buzzer-beater shot, or a specific play involving Magic Johnson or Steve Nash?⋯

Answer of human-simulating agent before aggregation at round 5:
- The image appears to be from an NBA All-Star Game, ⋯. The scoreboard shows EAST 74 and WEST 75 ⋯
- ⋯

Answer of human-simulating agent After aggregation at round 5:
The moment in question is from an NBA All-Star Game, as indicated by the \"ALL-STAR\" text on the court ⋯.

Reranked top 10 videos at round 5:
hbE29pZh76I_3_8, xkNDE1JL6Z8_50_59, nTUONeDqhdk_10_15, x8Ul09tk6zk_0_8, pzq5fPfsPZg_145_160, z0zb—BOhDY_16_22, fvBs0xpEZhQ_10_30, yBK02O1Oewc_20_35, xxHx6s_DbUo_173_177, pdrBPJYfTC8_33_39

Rank of target video: 1

Figure 5: Qualitative evaluation of MERLIN on MSVD. sample: hbE29pZh76I_3_8.

Case study for MSR-VTT

Initial query from dataset:
a man is performing live.

Top 10_candidates at round 0:
video8251, video7419, video8466, video7961, video8445, video7598, video7135, video7205, video7156, video7377

Rank of target video: 361

Anchor caption (Video caption of lc9bA-hvqHU_1_6):
A person wearing a red cap and sunglasses is playing an acoustic guitar and singing into a microphone. The guitar has ⋯

Generated question at round 1:
Can you describe any distinctive features or notable elements in the background of the video you have in mind, such⋯?

Answer of human-simulating agent before aggregation at round 1:
- The background of the video is quite minimalistic ⋯ The focus is primarily on the person in the foreground, who is ⋯
- In the image you provided, the background appears to be a plain, solid-colored wall with no distinctive features, ⋯
- ...

Answer of human-simulating agent After aggregation at round 1:
The background of the video is minimalistic and plain, featuring a solid, light-colored wall with no distinctive features ⋯

Reranked top 10 videos at round 1:
video8251, video7419, video7205, video8466, video7598, video7135, video8445, video7961, video7377, video7575

Rank of target video: 197

Anchor caption (Video caption of lc9bA-hvqHU_1_6):
Same as previous anchor caption.

Generated question at round 2:
Can you describe the action or activity the person in the video you have in mind is engaged in, particularly focusing ⋯ ?

...

Reranked top 10 videos at round 2:
video8251, video7419, video7205, video8466, video7135, video7598, video8445, video7377, video7558, video7575

Rank of target video: 14

Anchor caption (Video caption of hbE29pZh76I_3_8):
Same as previous anchor caption.

Generated question at round 3:
Can you describe the attire or appearance of the person in the video you have in mind, including any distinctive cloth⋯?

Answer of human-simulating agent before aggregation at round 3:
- The person in the image is wearing a sleeveless yellow dress. They have long, dark hair that is styled in a ponytail. ⋯
- In the image, the person is ⋯. They are holding a glass with a drink in it, which appears to be a wine glass. ⋯
- ⋯

Answer of human-simulating agent After aggregation at round 3:
The person in the video is wearing a sleeveless yellow dress. They have long, dark hair styled in a ponytail over one ⋯

Reranked top 10 videos at round 3:
video8471, video7419, video8251, video7205, video8466, video7135, video7726, video7598, video7412, video8445

Rank of target video: 1

...

Generated question at round 5:
Can you describe any specific actions or ⋯ such as whether they are smiling, gesturing, or interacting with anything ⋯ ?

Answer of human-simulating agent before aggregation at round 5:
- In the image, the person is holding a drink in their right hand and appears to be in the middle of speaking or making ⋯
- ⋯

Answer of human-simulating agent After aggregation at round 5:
The person in the video is holding a drink in their right hand and often gesturing with their left. They are wearing a ⋯.

Reranked top 10 videos at round 5:
video8471, video7726, video7419, video7915, video8251, video7725, video7116, video7216, video7412, video7573

Rank of target video: 1

Figure 6: Qualitative evaluation of MERLIN on MSR-VTT1ka. sample: video8471.