VRAgent: Self-Refining Agent for Zero-Shot Multimodal Video Retrieval

Supplementary Material

This supplementary material includes the following:

1. MM-MSRVTT Test Set

We create a new test set with more complex, multimodal queries from the original MSRVTT test videos. We first manually create a small set of five complex multimodal queries from the MSRVTT test videos chosen at random. We use this seed set as few-shot in-context examples for generating multimodal queries from other videos. The queries are created using GPT-4o [4], by providing it 25 frames from the video and its ASR transcript. The frames are composed into a 5x5 grid and passed to the LLM as one image. We prompt the model with guidelines on generating complex multimodal queries combining both the video frames and the ASR transcript, and also provide it with the in-context examples. The MM-MSRVTT test set consists of 500 queries. Examples include: *"conversation in a cozy cafe about professional growth"*, *"reporter addressing radicalization issues in front of storefronts"*.

2. TVR-1200 Test Set

For evaluating our method on multimodal video retrieval, we create a test set by adapting the TVR dataset [5]. TVR is designed for video-moment retrieval, consisting of 17.4k videos for training, 2.2k videos for validation, and 2.2k sequestered videos for testing. We derive our test set from the public validation set of TVR. Each video from the val set is annotated with multiple moments, resulting in a total of 10.9k query-moment pairs. Every query in TVR is also marked with a *query type* – as either visual, transcript, or joint – indicating the modality from which the query is derived. From the validation set, we choose 1200 queries, all from separate videos (randomly chosen from all the moments within a video). These are balanced across *query types*, with 400 queries of each type. Each query, along with its temporally trimmed moment from the original video, forms the TVR-1200 test set for multimodal video retrieval.

3. Re-ranking Details

As described in Sec. 3.1 in the main paper, the merging function involves a re-ranking step after aggregating individual tool instruction results into a ranked list $\tilde{S}_i$, for a tool instruction set $\mathcal{T}_i$. This re-ranking step reorders the top-k videos in this initial ranked list $\tilde{S}_i$ to obtain the final ranked list S_i. We achieve this by verifying that the top-k retrieved results are visually aligned with the visual tool instruction $\hat{q}_i^{t_j}$. This verification of alignment with visual tool instructions is done by processing the top-k videos in order, and moving it lower in the final ranked list S_i if it does not match the description in $\hat{q}_i^{t_j}$.

More specifically, for each VIS tool call with instruction $\hat{q}_i^{t_j}$, we use the frame with maximum similarity to the instruction to perform verification using a MLLM. We use LLaVA-1.6-7B [7] and prompt it to answer the following question: "`Does this image contain {instruction}? YES/NO`". This approach to re-ranking synergistically combines VLMs and MLLMs for improving retrieval.

4. Analysis of Self-Evaluation Score vs GT Rank

Figure 1 shows the correlation between the self-evaluation score and GT rank of different tool instruction sets for a given query. For each query, the plot shows 80 tool instruction sets, using the top-1 score for evaluation. Here, we empirically validate the choice of top-1 score for evaluation, by observing that increasing score leads to decrease in rank (left and middle columns in Fig. 1). Hence, using this score as optimization objective during self-refinement leads to increase in recall at K, by by better aligning the tool instruction sets with the original query. In cases where the queries are very ambiguous, *e.g.* *"a woman introducing someone"*, we observe (right column in Fig. 1) that the tool instruction sets with high score do not correlate with lower ranks. This is potentially due to numerous videos in the collection matching the query, however only one of those is a *ground-truth* match as annotated in the MSRVTT test set.

5. Details about VIS and ASR tools

VIS tool: The visual search tool (VIS) in VRAgent is based on either CLIP [8], BLIP-2 [6] or InternVideo2 [11] models. **CLIP/BLIP-2.** A given query is passed through the text encoder to obtain query features, and we extract frame features using the vision encoder. We compute cosine similarity between the normalized frame-level features and the

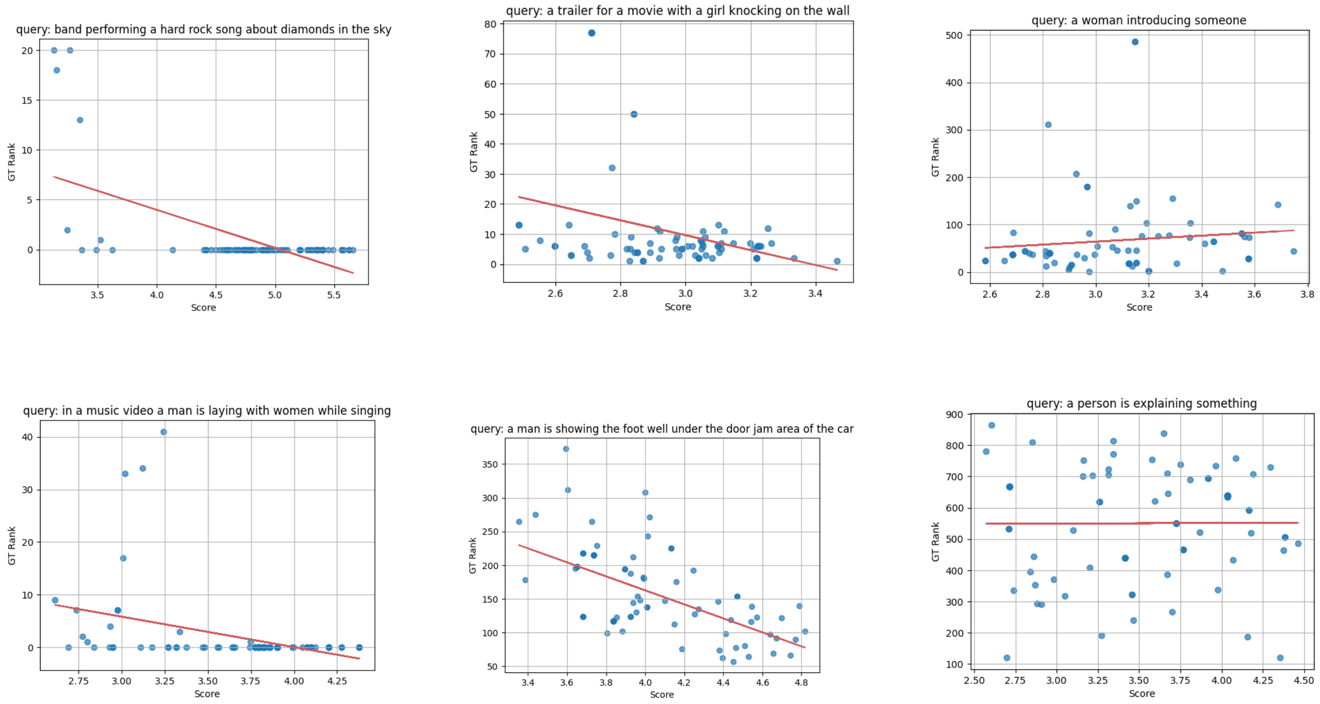

Figure 1. **Self-Evaluation Score vs GT Rank.** Increase in Top-1 score leads to decrease in GT Rank, validating the choice of our self-evaluation function. We notice that in cases with ambiguous original queries (right column), aligning the tool instruction set to the user intent is challenging. Such cases are good candidates for interactive mode for VRAgent, where we can obtain more context from the user via MCQs.

query feature, and the maximum cosine similarity over all frames (*max-sim*) is assigned as the visual search relevance score for the video-query pair. We choose *max-sim* because it effectively locates individual concepts/events, while the VRAgent decomposes the original user query into different concepts/events into tool instruction sets during self-refinement. **InternVideo2.** The video features in this case are obtained from the vision encoder using four frames from the video. We simply compute cosine similarity with the text feature of a query to get the relevance score for a given query-video pair.

ASR tool: The speech search tool from transcripts (ASR) in VRAgent is based on the Sentence-BERT model [9]. The S-BERT embeddings for each sentence in the transcript are averaged. We compute cosine similarity between this average transcript embedding and the S-BERT embedding of the query, and assign this as the transcript search relevance score for the video-query pair. We choose to mean-pool transcript embeddings since the video clips in TVR-1200 test set are of short durations.

6. Comparison with ViT-L/14

Here, we highlight that VRAgent using a smaller model in its VIS tool (ViT-B/16) can achieve better performance than using a larger model (ViT-L/14) without self-refinement (Tab. 1). The larger model only uses the original query with the VIS tool and does not perform self-refinement. This shows promise in improving the overall performance of the best foundation models even further through our agentic framework for video retrieval.

Method	R@1 ↑	R@5 ↑	R@10 ↑	AveR ↑
VIS tool (ViT-B/16)	31.7	56.1	67.0	51.6
VIS tool (ViT-L/14)	35.8	**57.4**	67.3	53.5
VRAgent (ViT-B/16)	**36.1**	57.0	**67.4**	**53.5**

Table 1. **Smaller model with VRAgent is better than larger model without VRAgent.** VIS tool baseline uses the original query for retrieval, whereas VRAgent translates and optimizes the original query into a tool instruction set, for aligning the user intent with the tool used.

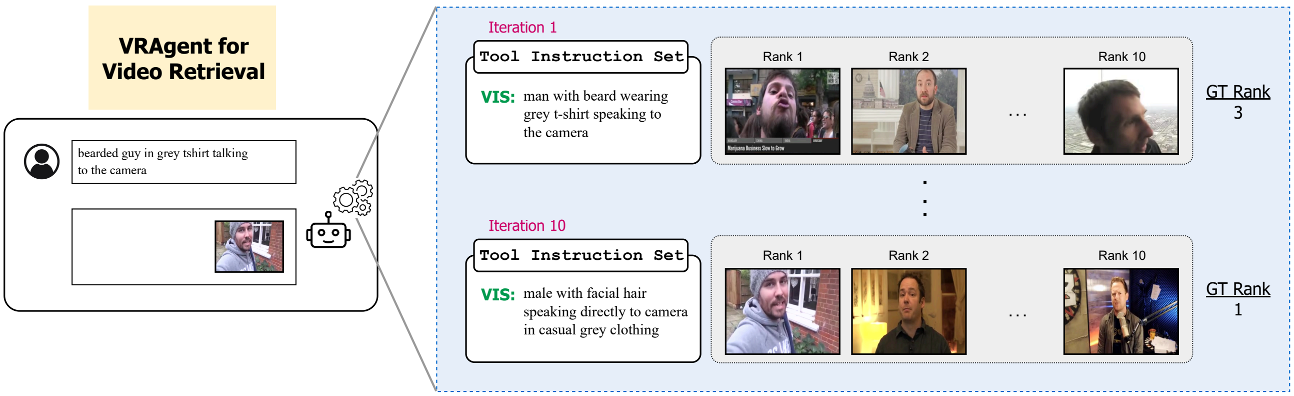
VRAgent for Video Retrieval
bearded guy in grey tshirt talking to the camera
Iteration 1
Tool Instruction Set
VIS: man with beard wearing grey t-shirt speaking to the camera
Rank 1
Rank 2
Rank 10
GT Rank 3
Iteration 10
Tool Instruction Set
VIS: male with facial hair speaking directly to camera in casual grey clothing
Rank 1
Rank 2
Rank 10
GT Rank 1

Figure 2. Qualitative example (from MSRVTT 1k-A test set) of VRAgent for text-to-video retrieval using only visual modality.

7. Computational time comparison

The computational time of VRAgent (iters=10, b=8) is 5 s per query compared to 0-shot evaluation of InternVideo2-1B which takes 0.41 s per query [10], while significantly improving the results – R@1 on MM-MSRVTT from 25.8 to 50.4, and R@1 on TVR-1200 from 14.8 to 20.2. Additionally, agentic approaches have become more mainstream in recent times. These are enabled by rapid improvements in LLM inference over recent years with model distillation, quantization and hardware acceleration combining to significantly reduce latency.

8. Additional Qualitative Examples

Figures 2 to 4 presents additional qualitative results for VRAgent for video retrieval and multimodal video retrieval respectively. In Fig. 2, we observe that the self-refinement procedure changes *"beard"* to *"facial hair"*, *"man"* to *"male"*, *etc*. to align the original user query with the visual tool. In Fig. 3, for multimodal video retrieval, VRAgent adds context-specific details in both the visual and speech modalities for improving the rank from 15 to 1. Fig. 4 shows an example for interactive video retrieval, where the VRAgent generates relevant MCQ questions with questions and options grounded in the video collection, allowing the user to provide additional context for retrieving the correct video, which improves the GT rank from 8 (iteration 1) to 1 (iteration 10).

9. Failure Cases

We analyzed the failure modes and find that VRAgent struggles in cases where the query lacks sufficient details. We design and incorporate interactive mode in VRAgent for such cases, making it possible to gather more context. The video foundation model we use extracts features based on 4 uniformly sampled frames from the video, which causes failures for some queries involving verbs or some fast moving actions.

10. Limitations

Retrieval typically involves two stages: indexing and querying. During the querying stage, our iterative self-refinement approach incurs a higher computational cost as compared with direct single-shot retrieval baselines. However, LLM inference has seen rapid improvements over recent years, with model distillation [1], quantization [3] and hardware accelerations [2] significantly reducing latency. As these trends continue, agentic workflows with test-time reasoning will progressively close the efficiency gap. For the indexing stage, which is performed once, our method has a computational cost similar to baseline approaches, requiring per-modality feature extraction.

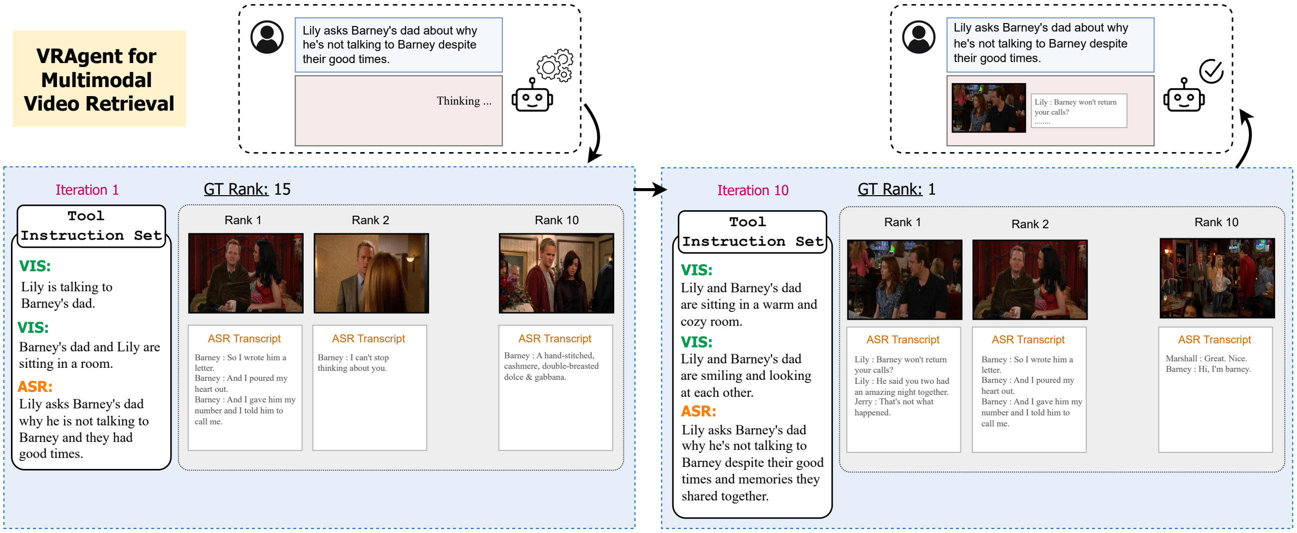
VRAgent for Multimodal Video Retrieval
Lily asks Barney's dad about why he's not talking to Barney despite their good times.
Thinking ...
Iteration 1
GT Rank: 15
Tool Instruction Set
VIS:
Lily is talking to Barney's dad.
VIS:
Barney's dad and Lily are sitting in a room.
ASR:
Lily asks Barney's dad why he is not talking to Barney and they had good times.
Rank 1
Rank 2
Rank 10
ASR Transcript
Barney : So I wrote him a letter.
Barney : And I poured my heart out.
Barney : And I gave him my number and I told him to call me.
ASR Transcript
Barney : I can't stop thinking about you.
ASR Transcript
Barney : A hand-stitched, cashmere, double-breasted dolce & gabbana.
Lily asks Barney's dad about why he's not talking to Barney despite their good times.
Lily : Barney won't return your calls?
....
Iteration 10
GT Rank: 1
Tool Instruction Set
VIS:
Lily and Barney's dad are sitting in a warm and cozy room.
VIS:
Lily and Barney's dad are smiling and looking at each other.
ASR:
Lily asks Barney's dad why he's not talking to Barney despite their good times and memories they shared together.
Rank 1
Rank 2
Rank 10
ASR Transcript
Lily : Barney won't return your calls?
Lily : He said you two had an amazing night together.
Jerry : That's not what happened.
ASR Transcript
Barney : So I wrote him a letter.
Barney : And I poured my heart out.
Barney : And I gave him my number and I told him to call me.
ASR Transcript
Marshall : Great. Nice.
Barney : Hi, I'm barney.

Figure 3. Qualitative example (from TVR-1200 test set) of VRAgent for multimodal video retrieval using visual and dialogue search tools.

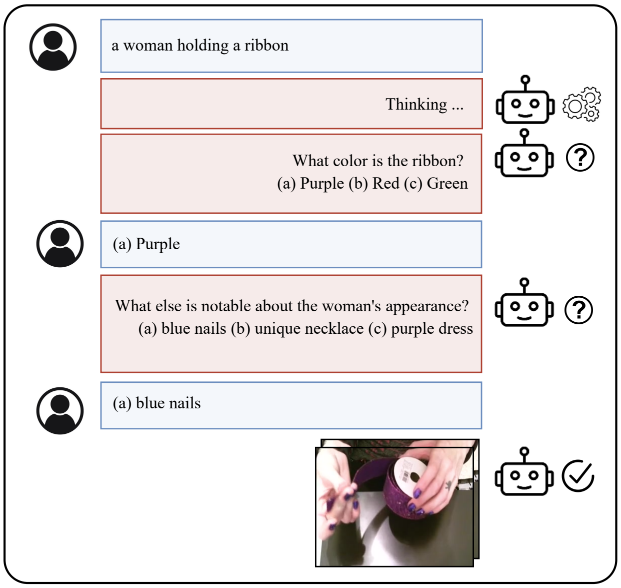
a woman holding a ribbon
Thinking ...
What color is the ribbon?
(a) Purple (b) Red (c) Green
(a) Purple
What else is notable about the woman's appearance?
(a) blue nails (b) unique necklace (c) purple dress
(a) blue nails

Figure 4. Qualitative example (from MSRVTT 1k-A test set) of VRAgent for interactive video retrieval using only visual search tool. The MCQ questions generated by VRAgent, which are grounded in the video collection, allows the user to provide additional context about the video.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 3

[2] Cerebras. Cerebras inference: 3× faster, 2025. Accessed: 2025-03-08. 3

[3] Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, et al. Apple intelligence foundation language models. *arXiv preprint arXiv:2407.21075*, 2024. 3

[4] 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. 1

[5] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal.

TVR: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020. 1

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[7] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. 1

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[9] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 2

[10] Jiapeng Wang, Chengyu Wang, Kunzhe Huang, Jun Huang, and Lianwen Jin. Videoclip-xl: Advancing long description understanding for video clip models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16061–16075, 2024. 3

[11] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 1