VideoSearch-R1: Iterative Video Retrieval and Reasoning via Soft Query Refinement

Seohyun Lee[1*], Seoung Choi[1*], Dohwan Ko[2*], Jongha Kim[2], and Hyunwoo J. Kim[1†]

[1] KAIST, Daejeon, Republic of Korea
{seohyunlee, choisw0823, hyunwoojkim}@kaist.ac.kr
[2] Korea University, Seoul, Republic of Korea
{ikodoh, jonghakim}@korea.ac.kr

Abstract. As video corpora continue to expand in both scale and task complexity, there is increasing demand for approaches that retrieve relevant videos from large-scale corpora (inter-video reasoning) and subsequently perform fine-grained, query-conditioned tasks (intra-video reasoning) within the retrieved content, such as temporal grounding. However, existing approaches typically treat retrieval as a preprocessing step, and consequently, when the initial retrieval fails, there is no mechanism to refine the search, leading to the failure of subsequent fine-grained intra-video reasoning. Moreover, while recent agentic frameworks have advanced video understanding, they typically assume that the query-relevant video is already given, focusing exclusively on intra-video reasoning tasks. To address these limitations, we propose **VideoSearch-R1**, an agentic framework for iterative video retrieval and reasoning through multi-turn interaction with a video search engine. Specifically, we introduce **Soft Query Refinement (SQR)** to refine search query tokens in a continuous latent space rather than rewriting queries in the discrete text space, enabling more efficient and fine-grained adjustments. SQR and its reasoning process are trained using Group Relative Policy Optimization (GRPO), guided by task-level reward signals derived from retrieval and downstream tasks. Building upon this, VideoSearch-R1 achieves state-of-the-art performance across three datasets on Video Corpus Moment Retrieval (VCMR), iteratively retrieving videos from large-scale corpora, refining search queries, and performing precise query-conditioned temporal grounding within the retrieved content. Our analyses show that SQR effectively refines the original query, requiring significantly fewer generated tokens than explicit text-level query refinement. Code and model checkpoints are publicly available at `mlvlab.github.io/VideoSearch-R1`.

Keywords: Reinforcement Learning · Agentic AI · Video Retrieval

1 Introduction

With the rapid growth of large-scale video corpora, recent studies have focused on efficiently and accurately retrieving relevant videos given a user query [8, 14,

[*] Equal contribution. [†] Corresponding author.

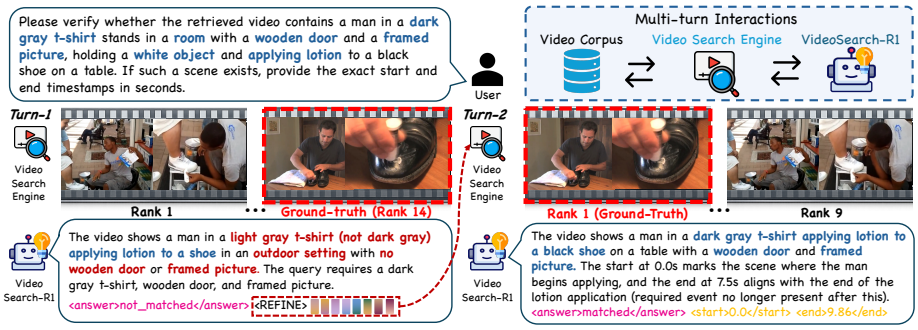


Fig. 1: An illustrative example of VideoSearch-R1. As an agentic AI system, VideoSearch-R1 enables multi-turn interaction through iterative video retrieval and reasoning, leveraging an external video search engine. This pipeline unifies corpus-level inter-video reasoning (*e.g.*, video retrieval) with intra-video reasoning (*e.g.*, temporal grounding) grounded in the retrieved video.

15,19–22,41,43]. Although these approaches achieve strong performance on standard video-level retrieval benchmarks through inter-video reasoning, identifying the correct video alone is insufficient for real-world applications. In practice, users require not only coarse inter-video reasoning but also query-specific intra-video reasoning within the retrieved video. For example, beyond identifying a relevant video, a system may need to conduct fine-grained reasoning, such as localizing the exact timestamp of a described event, extracting temporally grounded evidence, or performing question answering over the retrieved video.

However, existing pipelines [10, 47, 48] treat inter-video retrieval as a preprocessing stage prior to intra-video reasoning: inter-video retrieval models [8, 14, 15, 19–22, 41, 43] optimize for coarse-grained relevance over a video corpus, whereas intra-video reasoning modules [3, 5, 11, 35, 44, 45] operate independently within individual videos. Consequently, such pipelines are limited by their decoupled architecture, where failures in inter-video retrieval propagate to subsequent intra-video reasoning. These limitations motivate an *iterative video retrieval-and-reasoning* framework within an agentic AI system, where retrieval and reasoning are tightly integrated through an interactive loop. Rather than treating inter-video retrieval as a one-shot preprocessing step, the system autonomously retrieves relevant videos, dynamically refines search queries, and performs query-conditioned intra-video reasoning over the retrieved content. As in Fig. 1, this framework enables holistic inter-video reasoning across large-scale video corpora as well as fine-grained intra-video reasoning through multi-turn interaction.

Such agentic paradigms have recently shown strong potential in natural language processing through Retrieval-Augmented Generation (RAG), where search engines are treated as external tools [1, 13, 16]. For example, Search-R1 [13] introduces a reinforcement learning (RL)-based framework that enables Large Language Models (LLMs) to generate search queries and iteratively refine both queries and reasoning to provide a final answer. Inspired by these advances, similar efforts have emerged in video understanding, integrating search mech-

anisms with a core Vision-Language Model (VLM) [23, 29, 41]. In particular, video agentic frameworks incorporate external tools, such as object trackers, OCR models, temporal localizers, and video captioners, to facilitate long-form video understanding, where models often struggle to identify salient visual cues among extensive visual tokens. However, unlike text-based agentic systems that explicitly retrieve external knowledge, most existing video agentic frameworks implicitly assume that the query-relevant video is already known, *i.e.*, bypassing the video retrieval stage. While effective under this assumption, such designs become suboptimal when users expect the system to dynamically identify and retrieve relevant videos from large-scale corpora prior to intra-video reasoning.

To this end, we propose **VideoSearch-R1**, an agentic framework that integrates a video search engine to perform iterative video retrieval and reasoning through multi-turn interaction. The framework iteratively retrieves candidate videos, verifies query-video matching, refines search queries, and performs intra-video reasoning. For query refinement, instead of explicit text-level query refinement (*i.e.*, hard query refinement), we introduce **Soft Query Refinement (SQR)**, which generates query representations in the continuous latent space, enabling more efficient and fine-grained refinement. The soft query is appended to the original query to guide subsequent retrieval and is jointly optimized with the reasoning process via Group Relative Policy Optimization (GRPO) [30] to maximize task-level rewards. We train and evaluate VideoSearch-R1 on the Video Corpus Moment Retrieval (VCMR) [4] task, which requires the model to first retrieve the relevant video from a corpus given a textual query, and subsequently perform temporal grounding to predict the timestamp within the retrieved video that best corresponds to the query. VideoSearch-R1 achieves state-of-the-art performance across three VCMR benchmarks, ActivityNet-FIG, Charades-FIG, and DiDeMo-FIG, on both inter-video retrieval and intra-video temporal grounding. Our in-depth analysis shows that the proposed SQR effectively refines the original query to improve retrieval performance while requiring substantially fewer generated tokens than hard query refinement.

To summarize, our contributions are **threefold**:

- We propose VideoSearch-R1, an agentic framework for iterative video retrieval and reasoning. It iteratively retrieves candidate videos via a video search engine, verifies query-video matching, refines search queries, and performs intra-video reasoning through multi-turn interaction.
- We introduce Soft Query Refinement (SQR), which generates soft query tokens in a continuous latent space for fine-grained refinement, while requiring fewer generated tokens than hard query refinement.
- By jointly optimizing inter-video retrieval and intra-video reasoning within VideoSearch-R1, it achieves state-of-the-art performance on three VCMR benchmarks in both video retrieval and temporal grounding.

2 Related Works

Video retrieval and reranking. Video retrieval [8, 14, 15, 19–22, 41, 43] is a multi-modal task that retrieves the most relevant video given a text query. Early

approaches [8, 21, 22, 41, 43] commonly build upon CLIP [28] by encoding videos and texts with separate encoders and ranking candidates via cosine similarity in a shared embedding space. While efficient, this dual-encoder paradigm primarily captures coarse-grained alignment due to the lack of cross-modal interaction. To address this limitation, recent methods leverage VLMs [14, 15, 20, 36, 39] to model fine-grained cross-modal interactions and improve retrieval performance by reranking a fixed top-K set of candidate videos for each query.

Multi-turn reasoning of agentic AI. The multi-turn reasoning paradigm in agentic AI has recently attracted attention as an effective strategy for tackling complex analytical problems [12, 31, 37, 46]. In this setting, intermediate actions, such as tool invocation or document retrieval, are dynamically determined to progressively refine the solution. This framework has also demonstrated promise in video understanding, where multi-turn reasoning facilitates the interpretation of intricate temporal dependencies and long-range events [23, 24, 40]. Furthermore, recent works adopt RL methods, including GRPO [30], to enable optimization of reasoning trajectories and better align intermediate decisions with downstream objectives [13, 26, 27, 38, 49, 50]. In this work, we introduce VideoSearch-R1, which iteratively retrieves relevant videos, refines the query, verifies query-video alignment, and performs intra-video reasoning through multi-turn interactions.

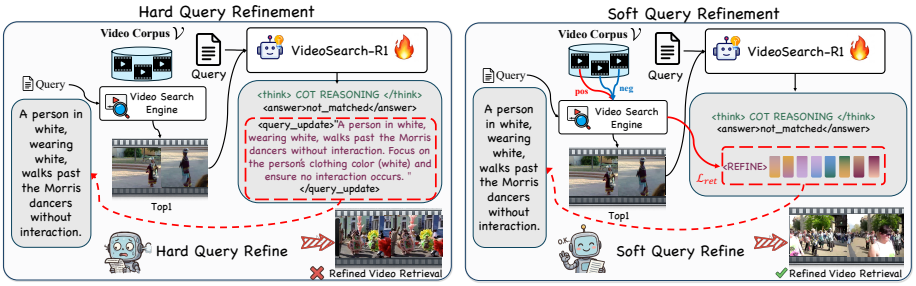
Soft reasoning in LLMs. Recent studies explore soft reasoning to reduce the overhead of explicit text-level chain-of-thought by updating continuous latent states to encode intermediate computations while minimizing long-form text generation [9, 17, 32–34, 42]. For instance, Coconut [9] leverages the final hidden state of an LLM as a compact representation of the reasoning state. Building upon this line of work, we extend the concept of soft reasoning to query refinement for enhanced video retrieval, and propose soft query refinement (SQR).

3 Method

We propose a video agentic model, **VideoSearch-R1**, an iterative video retrieval-and-reasoning framework that autonomously retrieves videos, verifies query-video matching, refines search queries, and performs intra-video reasoning. We also introduce **Soft Query Refinement (SQR)**, which produces query representations directly in the continuous latent space, enabling efficient and fine-grained adjustments. We first provide an overview of VideoSearch-R1, followed by a detailed description of its training and inference procedures.

3.1 VideoSearch-R1 with Soft Query Refinement

Given a user query, VideoSearch-R1 iteratively retrieves candidate videos through external search engine calls and verifies their relevance to the query. Let $t \geq 1$ denote the current turn of the iterative video retrieval-and-reasoning loop, initialized with the original user query q_1. At turn t, VideoSearch-R1 invokes a video search engine $\mathcal{R}$, a cross-modal dense embedding retriever (Qwen3-VL-Embedding-2B [18]), using the query q_t. The search engine returns the top-1



(a) Hard query refinement. **(b)** Soft query refinement (**Ours**).

Fig. 2: Comparison between hard query refinement and our Soft Query Refinement (SQR). SQR generates soft query tokens to perform fine-grained adjustments to the original query representation. In SQR, the soft query tokens are trained using the InfoNCE objective $\mathcal{L}_{\text{ret}}$, which provides richer discriminative supervision than the standard next-token prediction used in hard query refinement.

candidate video v_t from the video corpus $\mathcal{V}$ based on global similarity as:

$$v_t = \mathcal{R}(q_t) = \arg\max_{v \in \mathcal{V}} f(q_t)^\top f(v), \tag{1}$$

where f represents the encoder of the video search engine $\mathcal{R}$. However, video-level retrieval does not guarantee fine-grained semantic alignment with the query. To refine retrieval at turn t through iterative video retrieval and reasoning, VideoSearch-R1 performs a verification step with the reasoning process. Given the current query q_t and the retrieved video v_t, the model evaluates whether the video content satisfies the intended temporal semantics. It then generates a matching indicator $y_t^{\text{ret}} \in \{\texttt{'match'}, \texttt{'not match'}\}$, representing whether v_t aligns with q_t, along with the corresponding intermediate reasoning trace r_t.

If the retrieved result is deemed a mismatch (*i.e.*, $y_t^{\text{ret}} = \texttt{'not match'}$), the model performs SQR, which autoregressively generates a fixed number of soft query tokens $q_t^{\text{soft}} \in \mathbb{R}^{N \times D}$ as continuous latent embeddings, where N is the number of soft query tokens and D is the hidden dimension. Specifically, during autoregressive decoding, the hidden state corresponding to the previously generated token is projected through a linear layer and used directly as the input embedding for the next token. After generating N soft query tokens q_t^{soft}, we append them to the original query q_1 to form the refined query for the next turn, defined as $q_{t+1} = [q_1 \| q_t^{\text{soft}}]$. This refined query is then used to re-invoke the search engine. As in Fig. 2, unlike hard query refinement based on explicit text-level rewriting (Fig. 2a), SQR enables more fine-grained adjustments to the query representation while requiring fewer generated tokens (Fig. 2b).

This iterative retrieval and reasoning continues until the k-th turn, where either a valid match is identified or a predefined maximum number of turns T is reached. When a valid match is identified (*i.e.*, $y_t^{\text{ret}} = \texttt{'match'}$), the model proceeds to intra-video reasoning to predict the precise temporal boundaries y^{time} (*i.e.*, start and end timestamps) of the query-relevant segment within the

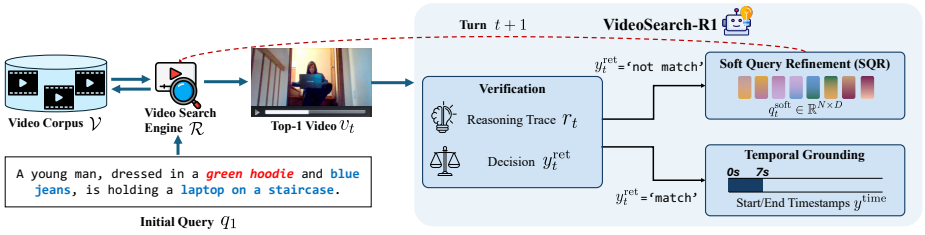


Fig. 3: Iterative video retrieval and reasoning of VideoSearch-R1. Given an initial query q_1, VideoSearch-R1 retrieves the top-1 video from a corpus via a video search engine and performs verification, producing a reasoning trace r_t and a matching decision y_t^{ret}. If $y_t^{\text{ret}} =$ `'not match'`, the model performs SQR by generating soft query tokens $q_t^{\text{soft}} \in \mathbb{R}^{N\times D}$ to construct a refined query $q_{t+1} = [q_1 \| q_t^{\text{soft}}]$. If matched, the model conducts temporal grounding to predict the start and end timestamps y^{time}.

retrieved video. If no valid match is found within the allowed iterations, the episode is treated as a failure case. The overall framework is illustrated in Fig. 3, and the template for each interaction turn is presented in Tab. 1.

3.2 Training Procedure

To optimize the iterative video retrieval-and-reasoning framework, we adopt a standard two-stage training pipeline widely used in video reasoning models [6, 7, 49]. In the first stage, we perform Supervised Fine-Tuning (SFT) to initialize VideoSearch-R1 with a structured reasoning template and encourage the generation of meaningful soft query tokens to improve video retrieval. The second stage applies RL-based policy optimization via GRPO [30], enabling the model to explore diverse reasoning trajectories and reinforce high-reward behaviors, thereby enhancing both inter-video retrieval accuracy and intra-video reasoning quality. We instantiate this training paradigm on the Video Corpus Moment Retrieval (VCMR) task, which naturally aligns with our unified objective: given a textual query, the model must first retrieve the relevant video from a large-scale corpus and subsequently predict the precise temporal boundaries within the retrieved video that best correspond to the query.

Stage 1: SFT cold start. In the SFT stage, we initialize VideoSearch-R1 with Qwen3-VL-2B-Instruct [2] and train it to follow a structured reasoning pattern while developing soft query generation capabilities within the iterative video retrieval-and-reasoning pipeline. Specifically, we supervise a reasoning trace r and two output variables, y^{ret} and y^{time}, corresponding to query-video matching verification and precise timestamp prediction, respectively. The reasoning trace and output variables are optimized using the following objectives:

$$\mathcal{L}_{\text{verif}} = -\log P(r, y^{\text{ret}} | q, v), \quad \mathcal{L}_{\text{time}} = -\log P(y^{\text{time}} | q, v, r, y^{\text{ret}}). \tag{2}$$

To obtain high-quality Chain-of-Thoughts (CoT) annotations of the reasoning trace r, we leverage a powerful VLM, Qwen3-VL-30B-A3B-Thinking [2]. We

Table 1: Template of a single turn within the multi-turn interaction of VideoSearch-R1. This process iterates until either the maximum number of turns is reached or the model verifies that the retrieved video matches the query.

System prompt: You are a video retrieval assistant. Your task is to analyze a retrieved video against the user query. Inside <think>...</think>, perform a step by step comparison between the query requirements and the visible evidence in the video. Identify whether a scene corresponding to the query appears in the video and determine the exact time span where it occurs. If a scene corresponding to the query appears in the video, output strictly in the following format: <think>...</think> <answer>matched</answer> <start>...</start> <end>...</end><REFINE>. Even if matched, you must still append the special token(s). <REFINE> at the very end to allow further latent refinement. If no scene corresponding to the query appears in the video, output strictly: <answer>not matched</answer><REFINE>. In this case, the special token(s) are required to initiate a latent query update. You must always append the special token(s) <REFINE> at the very end of the output. Do not invent details beyond what is visible. Be concise inside <think>...</think>. Do not output anything outside the specified tags.
User: [user query], [retrieved video]
Assistant (match): <think>...</think> <answer>matched</answer> <start>...</start> <end>...</end><REFINE>
Assistant (mismatch): <think>...</think> <answer>not matched</answer> <REFINE>

sample 2K query-video pairs from the VCMR dataset, consisting of 1K matching cases (where the search engine retrieves the correct video) and 1K negative cases (where retrieval fails). For positive (matching) query-video pairs, the model is trained to generate both an explanation of the semantic alignment between the query and the retrieved video, $\mathcal{L}_{\text{verif}}$, and a justification for the ground-truth temporal boundaries, $\mathcal{L}_{\text{time}}$. For negative pairs, supervision is applied only to $\mathcal{L}_{\text{verif}}$, encouraging the model to explain the semantic mismatch between the query and the video. Subsequently, the model is encouraged to refine the query via soft query generation and re-invoke the search engine in the next turn.

While CoT traces can be directly annotated from ground-truth query-video pairs, soft query tokens lack explicit supervision. To address this, we optimize soft queries using a contrastive objective based on InfoNCE [25] as in Fig. 2b. Concretely, after the model autoregressively generates N soft query tokens q^{soft}, these tokens are appended to the original query and optimized to maximize similarity with the ground-truth video v while minimizing similarity with negative videos $\mathcal{V}^{\text{neg}}$ in the search engine's embedding space, formulated as:

$$\mathcal{L}_{\text{ret}} = -\log\left(\frac{\exp\Big(f\big([q_1 \| q^{\text{soft}}]\big)^\top f(v)\Big)}{\exp\Big(f([q_1 \| q^{\text{soft}}])^\top f(v)\Big) + \sum_{v^- \in \mathcal{V}^{\text{neg}}} \exp\Big(f([q_1 \| q^{\text{soft}}])^\top f(v^-)\Big)}\right). \tag{3}$$

By explicitly incorporating negative video information through Eq. (3), this contrastive objective provides richer and more discriminative supervision for SQR than conventional next-token prediction used to train hard query refinement.

As a result, the overall SFT objective is defined as:

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{verif}} + \mathcal{L}_{\text{ret}} + \mathbb{1}_{y^{\text{ret}}=\text{`match'}}(\mathcal{L}_{\text{time}}). \tag{4}$$

Stage 2: Training via GRPO. Once the model acquires structured reasoning patterns and soft query generation capabilities, we proceed to the second stage to optimize the model using GRPO. While SFT provides a stable initialization, it does not explicitly optimize the interaction between retrieval and reasoning. We therefore employ RL to explore improved reasoning trajectories and more effective query refinement strategies. To align policy optimization with the objectives of the interleaved framework, we design four complementary reward signals: *format*, *verification*, *retrieval*, and *temporal grounding*.

The format reward R^{format} encourages the model to follow predefined reasoning and soft query structures, enforcing compliance with the template in Tab. 1, which reflects the iterative video retrieval and reasoning process. A reward of 1 is assigned if the model output strictly follows the required format, *e.g.*, the reasoning process enclosed in `<think>...</think>`, the matching result in `<answer>...</answer>`, and the predicted timestamps in `<start>...</start>` and `<end>...</end>`, and 0 otherwise. The verification reward R^{verif} supervises the query-video matching verification. A reward of 1 is assigned if the model correctly determines whether the video matches the query, and a reward of 0 is assigned otherwise. To supervise the soft query tokens during RL, we reuse the retrieval objective $\mathcal{L}_{\text{ret}}$ from the SFT stage and define the corresponding reward as $R^{\text{ret}} = \exp(-\mathcal{L}_{\text{ret}})$, encouraging improved retrieval performance. Finally, the temporal grounding reward R^{time} promotes accurate moment localization within the retrieved content by directly calculating the IoU between the predicted and ground-truth timestamps, reinforcing intra-video reasoning quality.

Overall, the final reward for the i-th sample is defined as:

$$R_i = R_i^{\text{format}} + R_i^{\text{verif}} + R_i^{\text{ret}} + \mathbb{1}_{y^{\text{ret}}=\text{`match'}}(R_i^{\text{time}}). \tag{5}$$

The advantage A_i is then computed by normalizing rewards within each group:

$$A_i = \frac{R_i - \text{mean}(\{R_j\})}{\text{std}(\{R_j\})}. \tag{6}$$

We adopt GRPO as the RL algorithm to train VideoSearch-R1, propagating reward signals across both inter-video retrieval and intra-video reasoning, resulting in a holistic optimization of iterative video retrieval and reasoning.

3.3 Inference via Multi-Turn Interaction

VideoSearch-R1 performs multi-turn interaction through iterative video retrieval and reasoning, consisting of external video search, retrieved video verification, soft query refinement, and temporal grounding within the selected video.

At each turn, the model first assesses the semantic alignment between the current query and the retrieved video, given the original query. If a mismatch is identified, it generates N soft query tokens, and these tokens are appended to the original query embedding sequence and fed into the video search engine $\mathcal{R}$. The search engine then re-ranks candidate videos conditioned on the refined query and returns the top-1 video. The newly retrieved video is incorporated into the context for the subsequent verification step. This multi-turn process continues until either the model verifies a correct match and produces a final temporal grounding prediction or the maximum number of retrieval turns T is reached, in which case the episode is considered a failure.

4 Experiments

Benchmarks and datasets. To evaluate the joint inter-video retrieval and intra-video reasoning capabilities of VideoSearch-R1, we adopt the recently introduced Video Corpus Moment Retrieval (VCMR) task [4], a challenging benchmark for corpus-level temporal grounding. VCMR requires the model to first retrieve the video relevant to a given textual query from the corpus (video retrieval) and then localize the precise start and end timestamps of the query within the retrieved video (temporal grounding). We conduct experiments on three VCMR benchmarks: ActivityNet-FIG, DiDeMo-FIG, and Charades-FIG.
Evaluation metrics. We report performance on VCMR and its subtask, video retrieval (VR). For VR, we adopt Recall@K (R@K) with $K \in \{1, 5, 10, 100\}$. For VCMR, we evaluate end-to-end performance using R@K under different temporal overlap thresholds, specifically IoU $\in \{0.3, 0.5, 0.7\}$ (denoted as IoU/R@1). A prediction is considered correct if the model identifies the ground-truth video and the IoU between the predicted and ground-truth temporal spans exceeds the specified threshold. Failure to retrieve the correct video yields a zero score for the VCMR metric. We additionally evaluate verification (VER) performance by measuring the accuracy of predicting `'match'` or `'not match'`.
Implementation details. We employ Qwen3-VL-Embedding-2B [18] as the video search engine, which retrieves the top-1 video given a textual query from a large-scale video corpus. VideoSearch-R1 is fine-tuned based on Qwen3-VL-2B-Instruct [2]. During training, the total number of visual tokens is set to 4,096 by sampling videos at 1 FPS with up to 64 frames. We use $N = 8$ soft query tokens. In Stage 1 (SFT), we fine-tune the model for 2K steps over 2K samples per dataset using AdamW with a learning rate of 2e-5. In Stage 2 (RL), we initialize from the SFT checkpoint and optimize with AdamW using a learning rate of 5e-7, weight decay of 0.01, and a maximum gradient norm of 1.0. We set the KL coefficient to $\beta = 0.01$, the rollout size to $G = 8$, and the decoding temperature to 1.0. We set the maximum number of inference turns to $T = 2$.
Baselines. To ensure a fair comparison, we introduce two baselines built upon the same backbone VLM, Qwen3-VL-2B-Instruct [2], with an identical number of parameters. The first baseline uses the zero-shot (ZS) model, while the second is fine-tuned (FT) on the same training dataset as VideoSearch-R1. The fine-

Table 2: Results on VCMR, VER, and VR.

Dataset	Method	VCMR			VER	VR			
		0.3/R@1	0.5/R@1	0.7/R@1	Acc	R@1	R@5	R@10	R@100
Charades-FIG	CONQUER [10]	–	1.2	0.7	–	2.8	9.0	14.1	51.7
	SQuiDNet [47]	–	2.6	0.9	–	11.7	33.9	44.0	51.7
	Qwen3-VL-2B (ZS)	12.2	7.2	2.9	30.0	21.6	41.8	51.5	84.2
	Qwen3-VL-2B (FT)	12.9	10.4	7.2	74.7				
	VideoSearch-R1	**16.5**	**13.4**	**8.2**	**75.7**	**24.6**	**42.9**	**52.2**	**84.7**
DiDeMo-FIG	CONQUER [10]	–	5.5	3.7	–	14.8	40.4	54.0	91.5
	SQuiDNet [47]	–	2.9	0.5	–	16.9	44.6	59.3	91.5
	Qwen3-VL-2B (ZS)	22.0	10.6	4.0	62.8	54.8	79.3	85.6	97.0
	Qwen3-VL-2B (FT)	23.6	22.1	16.7	73.1				
	VideoSearch-R1	**33.3**	**30.2**	**19.7**	**74.6**	**59.0**	**82.0**	**87.8**	**97.5**
ActivityNet-FIG	CONQUER [10]	–	3.0	1.6	–	13.5	36.4	50.0	89.3
	SQuiDNet [47]	–	4.7	2.1	–	32.6	79.9	87.9	89.3
	Qwen3-VL-2B (ZS)	17.2	10.1	5.8	63.0	55.1	78.8	86.7	98.1
	Qwen3-VL-2B (FT)	29.1	19.2	11.4	83.1				
	VideoSearch-R1	**33.8**	**22.3**	**12.3**	**83.3**	**61.1**	**81.7**	**88.5**	**98.4**

tuned baseline is trained to directly predict whether a retrieved video matches the query and to estimate its temporal boundaries, but it does not generate soft query tokens for refinement. We also apply the same multi-turn inference procedure as in VideoSearch-R1. Specifically, if the model determines that the top-1 retrieved video does not match the query, it sequentially evaluates the top-2, top-3, and subsequent candidates returned by the search engine.

4.1 Main Results

In Tab. 2, we evaluate VideoSearch-R1 on Charades-FIG, DiDeMo-FIG, and ActivityNet-FIG. Qwen3-VL-2B (ZS) and Qwen3-VL-2B (FT) achieve identical video retrieval (VR) results, as neither baseline updates the query representation during multi-turn inference. In contrast, VideoSearch-R1 iteratively refines the query representation via SQR, resulting in substantial improvements in VR despite using the same search engine. For example, on ActivityNet-FIG, R@1 improves by 6.0, underscoring the importance of iterative query refinement for accurate retrieval in multi-turn interaction. Furthermore, VideoSearch-R1 substantially outperforms both zero-shot and fine-tuned baselines in verification accuracy (VER) and temporal grounding for VCMR. Notably, VideoSearch-R1 improves 0.3/R@1 by 9.7 on DiDeMo-FIG compared to Qwen3-VL-2B (FT). These results demonstrate the effectiveness of jointly optimizing retrieval and reasoning within an agentic framework through RL-based policy learning.

4.2 Ablation Studies

Effect of training stages. In Tab. 3, compared to the zero-shot model, the SFT cold start (Stage 1) establishes a strong foundation by enforcing the prescribed

Table 3: Ablation studies of training stages on DiDeMo-FIG.

Method	VCMR			VER	VR			
	0.3/R@1	0.5/R@1	0.7/R@1	Acc	R@1	R@5	R@10	R@100
Qwen3-VL-2B (ZS)	22.0	10.6	4.0	62.8	54.8	79.3	85.6	97.0
VideoSearch-R1 (Stage1)	20.4	18.7	14.0	66.0	57.4	80.6	86.8	97.3
VideoSearch-R1 (Stage1 + Stage2)	**33.3**	**30.2**	**19.7**	**74.6**	**59.0**	**82.0**	**87.8**	**97.5**

Table 4: Ablation studies of reward design on DiDeMo-FIG.

Stage	Rewards			VCMR			VER	VR			
	R^{ret}	R^{verif}	R^{time}	0.3/R@1	0.5/R@1	0.7/R@1	Acc	R@1	R@5	R@10	R@100
Stage 1				20.4	18.7	14.0	66.0	57.4	80.6	86.8	97.3
Stage 1 + Stage 2	✔			18.6	17.3	13.0	65.0	58.0	81.4	87.4	97.5
Stage 1 + Stage 2	✔	✔		19.3	17.3	13.3	**75.0**	**59.7**	**82.6**	**88.3**	**97.7**
Stage 1 + Stage 2	✔	✔	✔	**33.3**	**30.2**	**19.7**	74.6	59.0	82.0	87.8	97.5

reasoning template and equipping the model with basic SQR capabilities, resulting in substantial improvements in R@1 for VR. However, the gains in temporal grounding on VCMR remain marginal. In contrast, Stage 2 (SFT + RL) yields pronounced improvements after applying GRPO. This discrepancy suggests that while SFT primarily enhances structural reasoning patterns, it is less effective at improving genuine temporal reasoning, which is further strengthened through subsequent RL training.

Ablation studies on reward design. To examine the individual contributions of each reward signal during the RL stage, we conduct an ablation study on R^{ret}, R^{verif}, and R^{time}, in Tab. 4. First, introducing the retrieval reward R^{ret} improves VR performance, suggesting that this signal encourages SQR to produce query representations that are better aligned with the corresponding video embeddings. Additionally, incorporating the verification reward R^{verif} improves the model's ability to assess semantic consistency between the retrieved video and the query, resulting in substantial improvements in verification accuracy and more reliable retrieval decisions. Finally, adding the temporal grounding reward R^{time} significantly improves temporal localization by promoting more precise reasoning and boundary prediction, with a slight trade-off in VER and VR. Overall, these results demonstrate that each reward component targets a distinct stage of the iterative retrieval-and-reasoning pipeline, and that their combination enables holistic optimization across structural formatting, retrieval alignment, verification reliability, and fine-grained temporal grounding.

4.3 Analysis

Analysis of SQR. We first present an in-depth analysis of SQR by examining how retrieval performance evolves as the number of soft query tokens increases. In Fig. 4, as additional soft query tokens are appended, the average R@1 consistently improves, underscoring that the soft query tokens incrementally refine

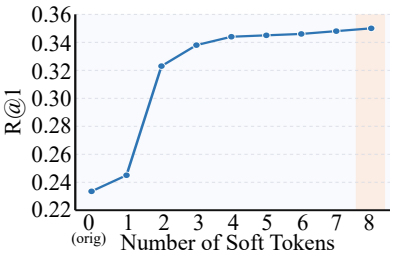


Fig. 4: Effect of the number of soft tokens. R@1 is computed over samples with refined queries.

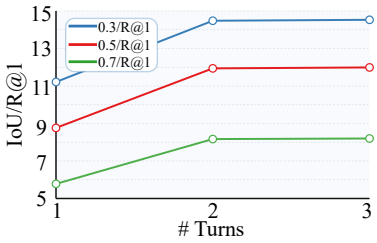


Fig. 5: Effect of multi-turn inference. The performance on VCMR saturates at $T = 3$.

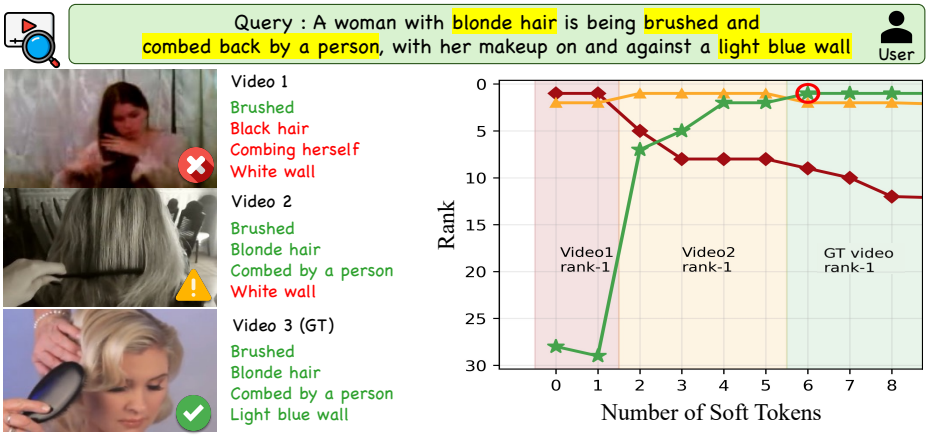


Fig. 6: Changes in the retrieved video as the number of soft tokens increases. The rank of the ground-truth video gradually improves as soft tokens are appended, capturing increasingly fine-grained semantics.

the original query. Fig. 6 illustrates a qualitative example where the rank of the ground-truth video gradually improves as soft tokens are appended to the original query. Without soft query tokens, the video search engine retrieves an incorrect video (video 1), capturing only the coarse concept 'a woman brushing'. As additional soft tokens are introduced, the retrieval results begin to capture more specific attributes, such as 'blonde hair' and 'combed by a person', in the video 2. Finally, with eight soft tokens, the search engine successfully retrieves the ground-truth video (video 3) at rank 1 by further capturing the context 'light blue wall'. These findings indicate that SQR effectively steers the continuous query representation toward the target video embedding, thereby improving retrieval accuracy and establishing a stronger foundation for intra-video temporal grounding.

To further analyze the necessity of continuous latent refinement, we compare SQR with hard query refinement (HQR) based on explicit text-level query

Table 5: Quantitative results of hard query refinement (HQR) and our soft query refinement (SQR) on ActivityNet-FIG

Method	VCMR			VER	VR				# tokens
	0.3/R@1	0.5/R@1	0.7/R@1	Acc	R@1	R@5	R@10	R@100	
Qwen3-VL-2B (ZS)	17.2	10.1	5.8	63.0	55.1	78.8	86.7	98.1	-
VideoSearch-R1 + HQR	33.2	**22.3**	11.9	82.2	57.6	77.2	85.4	97.8	26.8
VideoSearch-R1 + SQR	**33.8**	**22.3**	**12.3**	**83.3**	**61.1**	**81.7**	**88.5**	**98.4**	**8.0**

refinement. As shown in Tab. 5, SQR improves R@1 in VR by 7.2, compared to a 3.7 gain achieved by HQR. This indicates that directly optimizing continuous query representations enables more fine-grained adjustments than relying on discrete query refinement. Moreover, HQR produces substantially longer and more verbose refined queries (averaging 26.8 tokens), whereas SQR requires only eight soft query tokens to attain superior performance. Fig. 7 provides a qualitative example in which HQR retrieves an incorrect video, whereas SQR successfully refines the query representation and retrieves the correct video using only eight latent tokens. Notably, even after applying HQR, the search engine still returns the incorrect video at rank 1. We hypothesize that the longer rewritten queries produced by HQR introduce semantic noise, which can confuse the video search engine that relies on cross-modal embedding similarity rather than strong instruction-following capabilities. In contrast, SQR operates directly in the continuous embedding space, enabling more precise adjustments to the query representation using only a small number of latent tokens. Overall, these results demonstrate that SQR refines query embeddings more efficiently and precisely than HQR, leading to improved retrieval performance while requiring significantly fewer generated tokens.

Effect of multi-turn inference. Fig. 5 shows the performance trends as the number of inference turns increases. We observe a clear improvement from the first to the second turn, after which performance saturates when $T = 3$. This suggests that a small number of refinement turns is sufficient to effectively balance computational efficiency and retrieval accuracy.

Qualitative results. Finally, we present a qualitative case study in Fig. 1 that illustrates the multi-turn reasoning process of VideoSearch-R1. Given the textual query about 'a man in a dark gray t-shirt applying lotion to a black shoe indoors', the initial retrieval at $t = 1$ returns a rank-1 candidate video that shares superficial visual similarities (*e.g.*, a man applying lotion to a shoe) but fails to capture the detailed semantics. During verification, the model identifies this discrepancy and generates an intermediate reasoning trace explaining the missing semantic cues (*e.g.*, noting that the man wears a light gray t-shirt in an outdoor setting with no wooden door or framed picture), correctly predicting a `'not match'` decision followed by a `<REFINE>` token. Based on this mismatch, the model generates soft query tokens via SQR. When the search engine is re-invoked at $t = 2$, the refined query representation retrieves the ground-truth video at rank-1, which was previously ranked 14. After confirming the match

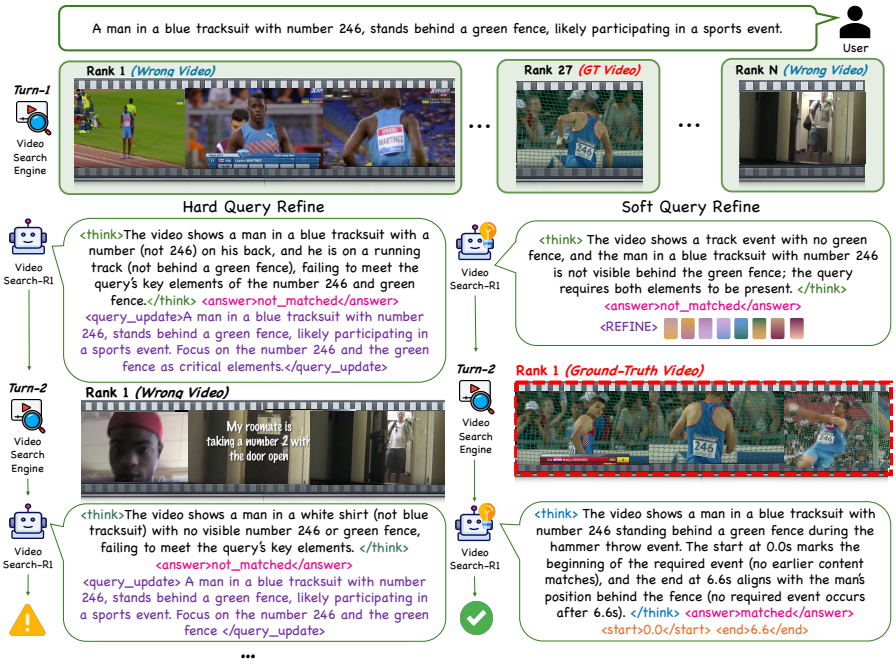


Fig. 7: Qualitative comparison between SQR and HQR.

(`'match'`), the model accurately localizes the target moment (start: 0.0s, end: 9.86s) with an IoU of 0.89. This example demonstrates the self-correcting capability of our iterative retrieval-and-reasoning framework, which resolves earlier retrieval errors through latent-space query refinement and subsequently performs fine-grained temporal reasoning.

5 Conclusion

In this paper, we introduce VideoSearch-R1, an agentic framework that unifies inter-video retrieval and intra-video reasoning within an iterative multi-turn loop. The model autonomously retrieves candidate videos, verifies their semantic alignment with user intent, refines search queries, and performs reasoning grounded in the retrieved content. We further propose Soft Query Refinement (SQR), a continuous latent-space query optimization mechanism that replaces explicit token-level rewriting. By avoiding verbose textual rewriting, SQR enables efficient and fine-grained query adjustments with substantially fewer generated tokens. Trained with reinforcement learning, VideoSearch-R1 achieves state-of-the-art performance on VCMR, demonstrating that the iterative retrieval-and-reasoning pipeline provides a robust, self-correcting foundation for complex, large-scale video understanding.

Acknowledgement

This work was supported by the InnoCORE program of the Ministry of Science and ICT (AI Meta-Scientist, N10260110, 30%), the Electronics and Telecommunications Research Institute (ETRI) grant funded by Korean government [26ZR1200, Research on Autonomous Vision Augmentation and Extension Technologies, 30%], and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00443251, Accurate and Safe Multimodal, Multilingual Personalized AI Tutors, 40%).

References

1. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-rag: Learning to retrieve, generate, and critique through self-reflection. In: ICLR (2024)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Cao, Z., Zhang, B., Du, H., Yu, X., Li, X., Wang, S.: Flashvtg: Feature layering and adaptive score handling network for video temporal grounding. In: WACV (2025)
4. Chen, H., Wang, X., Chen, H., Zhang, Z., Feng, W., Huang, B., Jia, J., Zhu, W.: Verified: A video corpus moment retrieval benchmark for fine-grained video understanding. In: NeurIPS (2024)
5. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
6. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. In: NeurIPS (2025)
7. Feng, K., Zhang, M., Li, H., Fan, K., Chen, S., Jiang, Y., Zheng, D., Sun, P., Zhang, Y., Sun, H., et al.: Onethinker: All-in-one reasoning model for image and video. arXiv preprint arXiv:2512.03043 (2025)
8. Gorti, S.K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A., Yu, G.: X-pool: Cross-modal language-video attention for text-video retrieval. In: CVPR (2022)
9. Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., Tian, Y.: Training large language models to reason in a continuous latent space. In: COLM (2025)
10. Hou, Z., Ngo, C.W., Chan, W.K.: Conquer: Contextual query-aware ranking for video corpus moment retrieval. In: ACM MM (2021)
11. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14271–14280 (2024)
12. Jain, C., Wu, Y., Zeng, Y., Liu, J., Dai, S., Shao, Z., Wu, Q., Wang, H.: Simpledoc: Multi-modal document understanding with dual-cue page retrieval and iterative refinement. In: EMNLP (2025)
13. Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H., Han, J.: Search-r1: Training llms to reason and leverage search engines with reinforcement learning. In: COLM (2025)
14. Ko, D., Lee, J.S., Choi, M., Meng, Z., Kim, H.J.: Bidirectional likelihood estimation with multi-modal large language models for text-video retrieval. In: ICCV (2025)

15. Lee, J.S., Ko, B., Cho, J., Lee, H., Byun, J., Kim, H.J.: Captioning for text-video retrieval via dual-group direct preference optimization. In: EMNLP Findings (2025)
16. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. NeurIPS (2020)
17. Li, B., Sun, X., Liu, J., Wang, Z., Wu, J., Yu, X., Chen, H., Barsoum, E., Chen, M., Liu, Z.: Latent visual reasoning. In: ICLR (2026)
18. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026)
19. Lin, S.C., Lee, C., Shoeybi, M., Lin, J., Catanzaro, B., Ping, W.: Mm-embed: Universal multimodal retrieval with multimodal llms. In: ICLR (2025)
20. Liu, Y., Zhang, Y., Cai, J., Jiang, X., Hu, Y., Yao, J., Wang, Y., Xie, W.: Lamra: Large multimodal model as your advanced retrieval assistant. In: CVPR (2025)
21. Liu, Y., Xiong, P., Xu, L., Cao, S., Jin, Q.: Ts2-net: Token shift and selection transformer for text-video retrieval. In: ECCV (2022)
22. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)
23. Luo, Y., Zheng, X., Li, G., Yin, S., Lin, H., Fu, C., Huang, J., Ji, J., Chao, F., Luo, J., et al.: Video-rag: Visually-aligned retrieval-augmented long video comprehension. In: NeurIPS (2025)
24. Min, J., Buch, S., Nagrani, A., Cho, M., Schmid, C.: Morevqa: Exploring modular reasoning models for video question answering. In: CVPR (2024)
25. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
26. Park, J., Na, J., Kim, J., Kim, H.J.: Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. In: NeurIPS (2025)
27. Peiyuan, F., He, Y., Huang, G., Lin, Y., Zhang, H., Zhang, Y., Li, H.: Agile: A novel reinforcement learning framework of llm agents. NeurIPS (2024)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
29. Ren, X., Xu, L., Xia, L., Wang, S., Yin, D., Huang, C.: Videorag: Retrieval-augmented generation with extreme long-context videos. arXiv preprint arXiv:2502.01549 (2025)
30. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
31. Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. NeurIPS (2023)
32. Su, D., Zhu, H., Xu, Y., Jiao, J., Tian, Y., Zheng, Q.: Token assorted: Mixing latent and text tokens for improved language model reasoning. In: ICML (2025)
33. Sun, G., Hua, H., Wang, J., Luo, J., Dianat, S., RABBANI, M., Rao, R., Tao, Z.: Latent chain-of-thought for visual reasoning. In: NeurIPS (2025)
34. Tan, W., Li, J., Ju, J., Luo, Z., Song, R., Luan, J.: Think silently, think fast: Dynamic latent compression of llm reasoning chains. In: NeurIPS (2025)
35. Wang, C., Feng, K., Chen, D., Wang, Z., Li, Z., Gao, S., Meng, M., Zhou, X., Zhang, M., Shang, Y., et al.: Adatooler-v: Adaptive tool-use for images and videos. arXiv preprint arXiv:2512.16918 (2025)

36. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
37. Wang, Q., Ding, R., Chen, Z., Wu, W., Wang, S., Xie, P., Zhao, F.: Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In: EMNLP (2025)
38. Wang, Q., Ding, R., Zeng, Y., Chen, Z., Chen, L., Wang, S., Xie, P., Huang, F., Zhao, F.: Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. In: NeurIPS (2025)
39. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: ECCV (2024)
40. Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In: CVPR (2025)
41. Wu, W., Luo, H., Fang, B., Wang, J., Ouyang, W.: Cap4video: What can auxiliary captions do for text-video retrieval? In: CVPR (2023)
42. Xu, Y., Guo, X., Zeng, Z., Miao, C.: Softcot: Soft chain-of-thought for efficient reasoning with llms. In: ACL (2025)
43. Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H., Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. arXiv preprint arXiv:2209.06430 (2022)
44. Yan, Z., Li, X., He, Y., Yue, Z., Zeng, X., Wang, Y., Qiao, Y., Wang, L., Wang, Y.: Videochat-r1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. In: NeurIPS (2025)
45. Yang, A., Nagrani, A., Seo, P.H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., Schmid, C.: Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In: CVPR (2023)
46. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: ICLR (2023)
47. Yoon, S., Hong, J.W., Yoon, E., Kim, D., Kim, J., Yoon, H.S., Yoo, C.D.: Selective query-guided debiasing for video corpus moment retrieval. In: ECCV (2022)
48. Zhang, H., Sun, A., Jing, W., Nan, G., Zhen, L., Zhou, J.T., Goh, R.S.M.: Video corpus moment retrieval with contrastive learning. In: SIGIR (2021)
49. Zhang, H., Gu, X., Li, J., Ma, C., Bai, S., Zhang, C., Zhang, B., Zhou, Z., He, D., Tang, Y.: Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In: CVPR (2026)
50. Zhou, Y., He, Y., Su, Y., Han, S., Jang, J., Bertasius, G., Bansal, M., Yao, H.: Reagent-v: A reward-driven multi-agent framework for video understanding. In: NeurIPS (2025)