RECOVR: CLOSING THE LOOP IN INTERACTIVE COMPOSED VIDEO RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Composed video retrieval (CoVR) retrieves a target video from a reference video and a modification text, yet a single query often leaves the target underspecified and existing methods offer no way to refine it. We formalize *interactive composed video retrieval* (ICoVR), a multi-turn extension of CoVR in which users refine their intent through natural-language feedback. Adapting interactive text-to-video retrieval methods to ICoVR exposes two structural weaknesses: they route all feedback through a *single retrieval channel*, and they operate *open-loop*, updating the query from feedback without checking whether their own rankings are drifting or stagnating. After a turn that fails to improve the target rank, the strongest adapted baseline improves on the next turn in only 30.12% of cases. We propose **ReCoVR**, a training-free framework that closes this loop by treating the system's own ranking trajectory as evidence alongside user feedback. An Intent Pathway routes each feedback turn to text-to-video and composed retrieval channels, and a Reflection Pathway accumulates constraints from feedback and uses ranking history to demote rejected or stagnant results. On WebVid-CoVR-Test, ReCoVR reaches 74.30% Recall at 1 (R@1) after one feedback turn and 90.22% after five, and its margin over the strongest baseline widens from 7.2 to 15.2 points. The gains carry over to Dense-WebVid-CoVR and FineCVR, and evaluation with real human feedback shows the same improvement trend.

1 INTRODUCTION

Composed video retrieval (CoVR) (Ventura et al., 2024a;b) searches for a target video using a reference video and a modification text, with steady progress in datasets (Ventura et al., 2024b; Thawakar et al., 2025; Yue et al., 2025) and methods (Thawakar et al., 2024; Yue et al., 2025). Existing approaches, however, assume a single query-response cycle. This assumption is restrictive for CoVR because a reference video and a short modification text often admit multiple plausible targets. Consider a user with a reference video of a city at night who asks the system to "`switch to daylight`." The system returns a daytime scene, but of a different city with dense skyscrapers instead of the low-rise layout in the reference. The user refines the query with "`keep a similar layout`," and the system now returns a matching layout at night, discarding the daylight constraint it had already satisfied. Under the single-turn paradigm (Fig. 1(a.1)), each refinement is a new query treated in isolation, so earlier progress can be undone without notice. In realistic visual search, users supply such missing constraints progressively after observing imperfect results (Marchionini, 2006), which motivates extending CoVR to multi-turn interaction.

Interactive retrieval has been explored in neighboring domains. In interactive text-to-video retrieval (I-TVR), prior methods typically ask clarifying questions, collect user answers, and use them to refine the query (Liang & Albanie, 2023; Han et al., 2024; Zhang et al., 2025). Related ideas also appear in visual dialog (Guo et al., 2018) and interactive composed image retrieval (Pal et al., 2023; Chen et al., 2025a; Luo et al., 2025). However, these settings do not address the composed video case, where the system must jointly maintain a visual reference and an evolving modification intent across turns. We therefore formulate **Interactive Composed Video Retrieval (ICoVR)**. The user starts with a standard CoVR query at Turn 0 and can refine the target through natural-language feedback from Turn 1 onward (Fig. 1(b)). This feedback can be spontaneous or answer an optional clarifying question from the system. To evaluate this setting, we adopt a controlled simulated-user protocol and adapt representative I-TVR methods as baselines.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

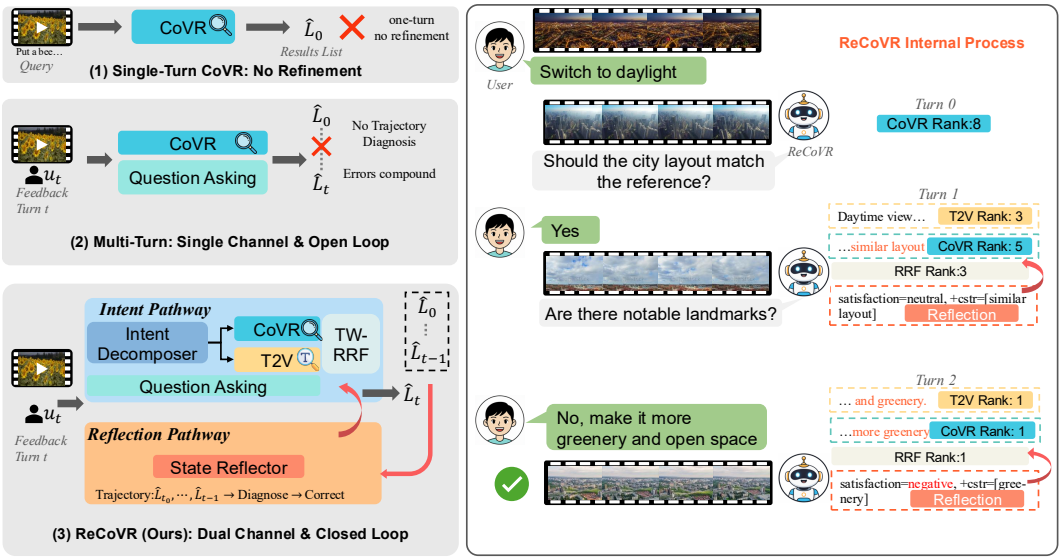

Figure 1: **Overview of Interactive Composed Video Retrieval (ICoVR) and ReCoVR. (a) Paradigm Comparison:** (1) Traditional CoVR is single-turn only. (2) Naive multi-turn methods use a single channel and an *open loop*: they update from user feedback without diagnosing past ranked results, allowing errors to compound. (3) Our framework **ReCoVR** uses an Intent Pathway that routes queries to composed video retrieval (CoVR) and text-to-video (T2V) channels, and a closed-loop Reflection Pathway that diagnoses past retrieval trajectories. **(b) Interaction Example:** Across turns, ReCoVR extracts constraints from feedback and ranks videos to reach the target.

Once transferred to ICoVR, these baselines reveal two structural weaknesses (Fig. 1(a.2)). First, they rely on a *single retrieval channel*. User feedback either describes a change to the current result or describes the target directly, yet a composed channel ties both forms to the reference video, even when a target description could be used directly by a text-to-video retriever. Second, they operate as *open-loop* systems: their query updates are driven by user feedback, but they do not check whether their own retrieval trajectory is drifting, stagnating, or repeatedly returning rejected candidates. This failure is measurable. The strongest adapted baseline fails to improve the target rank in 54.25% of its active feedback turns, and after such a turn it improves on the next executed turn in only 30.12% of cases (Sec. 4.2). Feedback alone thus often leaves earlier retrieval failures unresolved. Our key insight is that *the system's own ranking trajectory is an observable signal of these failures*: a top result that the user has rejected, or one that persists across turns, can be detected and corrected without any access to the target, using only rankings the system has already produced.

Building on this insight, we propose **ReCoVR** (**Re**flexive **Co**mposed **V**ideo **R**etrieval), a training-free closed-loop framework for ICoVR that reads its own retrieval trajectory alongside user feedback (Fig. 1(a.3)). An **Intent Pathway** addresses the single-channel limitation by decomposing each feedback turn into a direct target description and a relative edit, routing them to complementary text-to-video (T2V) and CoVR channels. A **Reflection Pathway** updates semantic constraints from user feedback and uses ranking history to detect repeated results. It then lowers the ranks of rejected or stagnant candidates before fusion. These pathways are supported by a dual memory bank that tracks semantic constraints and ranking history, and are unified by a time-weighted reciprocal rank fusion strategy that aggregates multi-turn evidence while emphasizing recent feedback. Because it requires no parameter updates, ReCoVR is backbone-agnostic and plugs into existing retrieval models. In summary, our contributions are three-fold:

1. We formulate **Interactive Composed Video Retrieval (ICoVR)**, a multi-turn extension of CoVR for progressive intent refinement through natural-language feedback, and identify single-channel execution and open-loop updating as weaknesses of adapted retrievers.

2. We propose **ReCoVR**, a training-free framework that routes feedback to complementary retrieval channels and uses the ranking trajectory to correct both the next query, via accumulated constraints, and the current ranking, via rank-level demotion.
3. ReCoVR outperforms adapted interactive baselines on three CoVR benchmarks, with an R@1 margin on WebVid-CoVR-Test that widens from 7.2 to 15.2 points over five turns. After a turn that fails to improve the target rank, it improves the rank on the next turn in 88.15% of cases versus 30.12% for the strongest baseline, and matched ablations attribute the gains to dual-channel routing, semantic reflection, and rank correction.

2 Related Work

Composed Video Retrieval. Composed Video Retrieval (CoVR) extends Composed Image Retrieval (Han et al., 2017; Vo et al., 2019; Chen & Bazzani, 2020; Liu et al., 2021; Wu et al., 2021) to the video domain, where a target video is retrieved from a reference video and a modification text. Existing methods have advanced through large-scale caption-derived training data (Ventura et al., 2024a;b), visual-bias mitigation by separating preserved content from text-driven changes (Thawakar et al., 2024; Chen et al., 2025b; Wang et al., 2025; Hu et al., 2026), and fine-grained temporal benchmarks or dense descriptions (Hummel et al., 2024; Gupta et al., 2025; Yue et al., 2025; Thawakar et al., 2025; Ji et al., 2026). The reason-aware CoVR-R (Thawakar et al., 2026) predicts edit after-effects such as state transitions, action phases, and camera changes, but still retrieves from a single fixed reference–edit pair. ICoVR additionally requires correcting drift, stagnation, and repeated feedback violations across turns.

Interactive Visual Retrieval. Interactive retrieval refines search through iterative user feedback. In the image domain, relevance feedback (Rui et al., 1998; Wu et al., 2004) evolved into state aggregation and sequence modeling (Guo et al., 2018; Cai et al., 2021; Yuan & Lam, 2021; Pal et al., 2023; Wei et al., 2023) and, recently, multi-turn reasoning with large multimodal models (Levy et al., 2023; Lee et al., 2024; Chen et al., 2025a; Feng et al., 2025; Lu et al., 2025; Bai et al., 2025b; He et al., 2025; Qin et al., 2025). For videos, interactive text-to-video retrieval (I-TVR) (Han et al., 2025) refines queries by appending answers (Liang & Albanie, 2023), interpolating embeddings (Han et al., 2024), rewriting dialogue history (Wu et al., 2024) or quantifying uncertainties (Zhang et al., 2025). None of these methods addresses the composed video setting, which must jointly maintain a visual reference and an evolving modification text. Beyond interactive retrieval, negative relevance feedback demotes documents a user has rejected (Wang et al., 2008; Bi et al., 2019), and self-reflective LLM agents revise later actions from earlier outcomes (Yao et al., 2022; Shinn et al., 2023). ReCoVR applies both ideas to interactive retrieval but reflects on a specific signal, the ranked-list trajectory of its own retrieval channels, and acts only through query constraints and rank-level corrections. Adapted I-TVR methods consume feedback but do not diagnose their past rankings.

3 Method

3.1 Problem Setup and ReCoVR Overview

Interactive Composed Video Retrieval (ICoVR) extends single-turn composed video retrieval (Ventura et al., 2024a;b) to multiple turns (Fig. 2(a)). Given a reference video v_{ref}, an initial modification text u_0, and a gallery $\mathcal{G}$, the user first issues a standard CoVR query at Turn 0 and receives an initial ranked list $\hat{L}_0$. When the target is not found, the user can continue the search from Turn 1 onward by providing natural-language feedback:

$$\{(q_t,\ u_t,\ \hat{L}_t)\}_{t=1}^{T}, \tag{1}$$

where T is the maximum number of feedback turns, q_t is an optional system-generated clarifying question, u_t is the user's feedback or refinement instruction, and $\hat{L}_t$ is the updated ranked list. The goal is to progressively align the ranked list with the user's intent over turns. The system never observes the target. When producing $\hat{L}_t$, its only information is the interaction history

$$\mathcal{H}_t = \big(v_{\text{ref}},\ u_0,\ \{(q_i, u_i)\}_{i=1}^{t},\ \{\hat{L}_i\}_{i=0}^{t-1}\big).$$

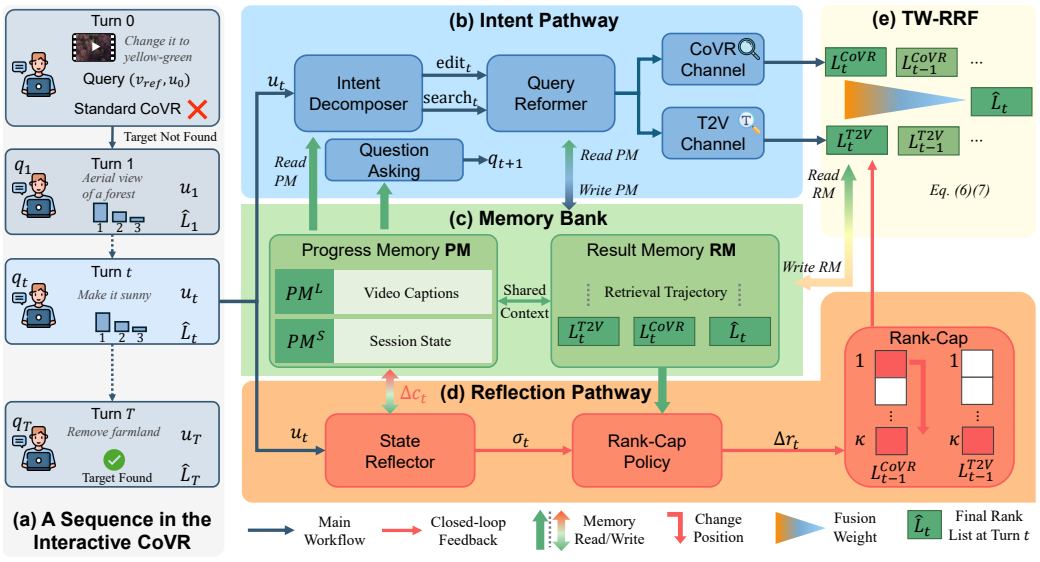

Figure 2: **Pipeline of ReCoVR.** When standard CoVR fails at Turn 0 **(a)**, the system enters an interactive loop. At each turn, the Intent Pathway **(b)** decomposes user feedback u_t into dual-channel queries and generates a clarifying question q_{t+1}, while the Reflection Pathway **(d)** infers satisfaction σ_t and constraints Δc_t from the session state, then consults the retrieval trajectory to apply rank caps Δr_t. Both pathways read from and write to a shared Memory Bank **(c)** that tracks semantic state (PM) and ranked results (RM). Finally, TW-RRF **(e)** fuses candidates over a recency-weighted sliding window into the final ranked list $\hat{L}_t$.

Existing interactive retrievers update their queries mainly from the dialogue $\{(q_i, u_i)\}$. ReCoVR is *closed-loop*: it also uses the ranking trajectory $\{\hat{L}_i\}_{i<t}$ to decide which candidates to demote and which result anchors the next composed query.

As shown in Fig. 2, each ReCoVR turn consists of three stages. The **Intent Pathway** decomposes feedback u_t into structured, channel-specific queries for two fixed retrieval engines: a text-to-video (T2V) channel and a composed video retrieval (CoVR) channel. The **Reflection Pathway** uses feedback and semantic state to update query constraints. It also uses ranking history to lower the ranks of rejected or stagnant candidates. Finally, **Time-Weighted Reciprocal Rank Fusion (TW-RRF)** aggregates candidates from both channels over recent turns to produce the final ranked list $\hat{L}_t$. ReCoVR uses two memory banks. **Progress Memory (PM)** stores semantic information. Its long-term part PM^L caches gallery captions and metadata for reuse across sessions. Its short-term part PM^S stores the running query and accumulated constraints for the current session. **Result Memory (RM)** stores ranked lists from both channels across turns. A vision-language model (VLM) supplies structured queries and correction signals rather than scoring videos directly, so ReCoVR runs on pre-trained retrievers without fine-tuning. Appendix F provides the turn-level algorithm. Appendix G lists the prompts and output schemas used by every reasoning module.

3.2 Reflexive Retrieval Mechanism

The Intent Pathway (Sec. 3.2.1) decides *what to search for* from the latest feedback. The Reflection Pathway (Sec. 3.2.2) decides *what to correct*: its constraints shape the next turn's queries, and rank-cap adjusts the current turn's lists before fusion. Both read the same session snapshot and can run in parallel; Algorithm 2 in Appendix F gives the full order of operations in each turn.

3.2.1 Intent Pathway

The Intent Pathway translates ambiguous user feedback into executable queries (Fig. 2(b)). In multi-turn retrieval, user feedback often conveys mixed instructions. For instance, "*Make the keyboard black, add backlighting*" mixes a relative edit with a target description; routing it to a single query

type would discard part of the signal, so the **Intent Decomposer** maps u_t into two queries:

$$(\text{search}_t,\ \text{edit}_t)\ =\ \mathcal{I}(u_t,\ PM^S_{t-1},\ PM^L). \tag{2}$$

Here, search_t is a target description synthesized from user feedback and accumulated context, while edit_t is a relative modification to the current result and may be empty for description-only feedback.

The **Query Reformer** appends the new intent to the running query in PM^S_{t-1}. It uses a large language model (LLM) to summarize text that exceeds a length threshold. This memory also contains constraints Δc_{t-1} from the previous turn (with $\Delta c_0 = \emptyset$), so the new query retains earlier corrections. Retrieval then proceeds through two channels:

$$L^{T2V}_t,\ L^{CoVR}_t\ =\ \mathcal{R}(\text{search}_t,\ \text{edit}_t,\ \hat{L}_{t-1}(1),\ \mathcal{G}). \tag{3}$$

Here, $\hat{L}_{t-1}(1)$ denotes the top result shown to the user at the previous turn. The **T2V Channel** encodes search_t and retrieves videos by text-to-video similarity. The **CoVR Channel** uses edit_t to change the previously displayed result. When the edit is non-empty, it uses $\hat{L}_{t-1}(1)$ as the visual reference and retrieves by composed similarity. This dynamic anchor links a user's edit to the video they just saw. T2V thus supports direct target descriptions, while CoVR supports relative changes to the displayed video, so each form of feedback reaches the retriever suited to it.

An optional **Question Asking** module reads PM^S_{t-1} and generates a clarifying question q_{t+1}. Prior I-TVR methods use questions and answers to refine queries (Liang & Albanie, 2023; Han et al., 2024; Zhang et al., 2025). ReCoVR also accepts direct descriptions and edit instructions: users can answer the question or provide other feedback. Answering is never required: when the simulated user ignores the question entirely and follows a fixed modify-only or rewrite-only policy, ReCoVR still reaches 86.15% and 92.14% R@1 at Turn 5 (Appendix D.4).

3.2.2 Reflection Pathway

The Reflection Pathway (Fig. 2(d)) combines semantic correction by the **State Reflector** with rank correction by the Rank-Cap Policy. The State Reflector evaluates user feedback u_t in the context of the session state stored in PM^S_{t-1}, including accumulated queries, prior constraints, and gallery captions from PM^L. This joint analysis generates two structured outputs:

$$(\sigma_t,\ \Delta c_t)\ =\ \mathcal{S}(u_t,\ PM^S_{t-1},\ PM^L). \tag{4}$$

We call this semantic update *State Reflection*. The satisfaction level $\sigma_t \in \{\text{negative}, \text{neutral}\}$ indicates whether the user explicitly rejects the current result or is providing incremental guidance. The constraint delta Δc_t extracts semantic conditions based on the gap between user expectations and the current query state. It derives positive constraints that accumulate to guide ongoing searches, alongside negative constraints that act as semantic filters during query rewriting.

Beyond semantic constraints, ReCoVR needs an immediate mechanism to prevent clearly unhelpful results from dominating the next interaction. Inspired by negative-feedback reranking (Wang et al., 2008; Bi et al., 2019), we instantiate a lightweight, failure-triggered **Rank-Cap Policy**:

$$\Delta r_t\ =\ \mathcal{P}(\sigma_t,\ RM_{t-1}). \tag{5}$$

Let $\hat{v} = \hat{L}_{t-1}(1)$ be the previously displayed top result. The policy triggers on *explicit rejection* or on *strategy stagnation*, where the displayed top result persists across turns, and selects a cap position

$$\kappa_t = \begin{cases} \kappa_{\text{neg}}, & \sigma_t = \text{negative}, \\ \kappa_{\text{neu}}, & \sigma_t = \text{neutral and } \hat{L}_{t-1}(1) = \hat{L}_{t-2}(1), \\ \infty, & \text{otherwise}, \end{cases}$$

with $\kappa_{\text{neu}} < \kappa_{\text{neg}}$. In each channel list $L^{(s)}_t$, rank-cap moves $\hat{v}$ to rank $\max(r^{(s)}_{t,\hat{v}}, \kappa_t)$ and preserves the relative order of all other candidates; $\kappa_t = \infty$ leaves the lists unchanged. Capping rather than removing $\hat{v}$ keeps a partially relevant result available to later turns, and acting on ranks avoids comparing similarity scores that are not calibrated across channels (Appendix A.1); stagnation receives the milder cap because a persisting result is weaker evidence than an explicit rejection. Demoting $\hat{v}$ also keeps an unreliable result from becoming the visual anchor of the next turn.

3.3 Time-Weighted Reciprocal Rank Fusion

As interaction progresses, ReCoVR collects ranked lists from both T2V and CoVR channels across turns. Standard Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) combines multiple ranked lists by rewarding candidates that appear consistently at high ranks. However, directly applying RRF to all historical turns ignores the temporal nature of ICoVR: later feedback usually reflects a more refined intent, while early turns may encode ambiguous or outdated constraints. We therefore extend RRF into a temporally aware fusion rule by adding recency weighting and a sliding temporal window, yielding **Time-Weighted Reciprocal Rank Fusion (TW-RRF)**.

For current turn t and window size W, TW-RRF aggregates $N = t - t_0 + 1$ turns, from $t_0 = \max(1, t - W + 1)$ to t. The normalized linear recency weight for turn t' is:

$$w_{t'} = \frac{2(t' - t_0 + 1)}{N(N+1)}. \tag{6}$$

Recent turns thus contribute more, and the window keeps early ambiguous feedback from dominating after the user has refined the intent. For a candidate video v_i, let $r^{(s)}_{t',i}$ be its rank in channel $s \in \{\text{T2V}, \text{CoVR}\}$ at turn t', and let $\mathbb{1}[v_i \in L^{(s)}_{t'}]$ indicate whether it appears in the top-K list. The final list $\hat{L}_t$ is obtained by sorting candidates according to:

$$\hat{L}_t = \underset{v_i}{\text{sort}^{\downarrow}} \left[\sum_{s \in \{\text{T2V},\text{CoVR}\}} \sum_{t'=t_0}^{t} w_{t'} \cdot \frac{\mathbb{1}[v_i \in L^{(s)}_{t'}]}{k + r^{(s)}_{t',i}} \right], \tag{7}$$

where k is the standard RRF smoothing constant. When all turns are assigned equal weights and the window covers the full history, TW-RRF reduces to conventional RRF over multi-turn channel outputs. TW-RRF operates on the channel lists after rank-cap (Eq. (5)): the current lists are adjusted before being stored in RM and fused, earlier lists are not rewritten, and an adjusted list influences later turns until it leaves the window (Appendix F).

4 Experiments

4.1 Experimental Setup

Datasets and Interactive Baselines. We evaluate on **WebVid-CoVR** (Ventura et al., 2024a) (2,556 manually annotated triplets), **Dense-WebVid-CoVR** (Thawakar et al., 2025) (the same video pairs with richer modification texts: 31.16 vs. 4.6 words on average), and **FineCVR** (Yue et al., 2025) (10,043 fine-grained temporal queries over a larger, more diverse gallery). Generalization benchmarks include **FashionIQ** (Wu et al., 2021) for composed image retrieval (three clothing categories), and the text-to-video benchmarks **MSRVTT-1kA** (Xu et al., 2016; Yu et al., 2018), and **AVSD-1k-Test** (Alamri et al., 2019) (1,000 queries each for the latter two, following prior protocols (Lyu et al., 2023; Madasu et al., 2022; Zhang et al., 2025)). We adapt `IVR` (Liang & Albanie, 2023), `Merlin` (Han et al., 2024), and `UMIVR` (Zhang et al., 2025) as interactive baselines, equipping each with the same single-turn CoVR retriever as ReCoVR.

User Simulator. Following simulated-user evaluation in interactive retrieval (Zhang et al., 2024; Balog & Zhai, 2023; Liang & Albanie, 2023; Zhang et al., 2025; Chen et al., 2026; Dou et al., 2025), a VLM compares the target video with the current top-ranked candidate to generate feedback. It serves solely as an *evaluation instrument*: ReCoVR and the baselines never access the target, receiving only feedback and their own retrieval history. All methods share the simulator, prompts, targets, initial rankings, and stopping criteria. The simulator chooses between a relative edit and a fresh target description from how far the displayed result is from the target, independently of the method under evaluation, and replacing ReCoVR's reasoning VLM with InternVL3.5 while keeping the simulator fixed preserves the gains (Appendix B). Appendix D details the simulator and its prompts, and Sec. 4.4 reports evaluation with real human feedback.

Evaluation Protocol. Starting from a CoVR query at Turn 0, each interaction allows up to $T{=}5$ feedback turns, presenting the top-ranked result before each update. We report per-turn Recall at K (R@K) and Best log Rank Integral (BRI) (Lee et al., 2024). BRI integrates the logarithm of the best target rank attained so far, rewarding early retrieval success across the trajectory; lower is better.

Table 1: **Comparison with interactive baselines** on WebVid-CoVR-Test. ReCoVR leads in R@1 at every feedback turn, with the margin over the strongest baseline growing from 7.2 to 15.2 points, and it matches or exceeds all baselines on R@5, R@10, and BRI at every turn.

	Methods	*Turn 0*	*Turn 1*	*Turn 2*	*Turn 3*	*Turn 4*	*Turn 5*
R@1 ↑	IVR	54.30	66.55	70.11	71.95	73.00	73.51
	Merlin	54.30	60.80	63.34	64.63	65.49	65.77
	UMIVR	54.30	67.10	71.09	72.85	74.10	75.04
	ReCoVR(ours)	54.30	**74.30**	**80.32**	**84.39**	**87.72**	**90.22**
R@5 ↑	IVR	77.00	85.45	87.68	88.93	89.75	90.10
	Merlin	77.00	79.23	82.75	83.92	84.43	84.62
	UMIVR	77.00	86.62	**89.28**	90.02	90.18	90.41
	ReCoVR(ours)	77.00	**87.28**	**89.28**	**91.28**	**93.04**	**94.13**
R@10 ↑	IVR	84.35	91.08	93.78	94.44	94.84	95.03
	Merlin	84.35	86.27	89.63	90.49	91.20	91.47
	UMIVR	84.35	92.64	93.56	94.33	94.56	94.99
	ReCoVR(ours)	84.35	**93.08**	**93.81**	**94.56**	**95.27**	**96.24**
BRI ↓	IVR	-	0.7368	0.6187	0.5560	0.5161	0.4880
	Merlin	-	0.8165	0.7399	0.6921	0.6603	0.6381
	UMIVR	-	0.7285	0.6049	0.5394	0.4986	0.4700
	ReCoVR(ours)	-	**0.6737**	**0.5271**	**0.4538**	**0.4034**	**0.3643**

Table 2: **Reference comparison with single-turn CoVR methods** on WebVid-CoVR-Test. Interactive turns use extra feedback, so this is not a like-for-like benchmark. **Bold** marks the first turn at which ReCoVR matches the best single-turn result.

Methods \ Metrics	**R@1** ↑	**R@5** ↑	**R@10** ↑	**R@50** ↑	**Avg.** ↑
Single-Turn CoVR					
CoVR$_{AAAI\ 24}$	53.13	79.93	86.85	97.69	79.40
ECDE$_{CVPR\ 24}$	60.12	84.32	91.27	98.72	83.61
CoVR-2$_{TPAMI\ 24}$	59.82	83.84	91.28	98.24	83.30
TFR-CVR$_{ECCV\ 24}$	51.70	75.30	80.70	-	-
FDCA$_{ICLR\ 25}$	54.80	82.27	89.84	97.70	81.15
HUD$_{MM\ 25}$	63.38	86.93	92.29	98.76	85.34
CADP$_{PR\ 25}$	63.62	87.36	93.19	99.30	85.87
REFINE$_{TOMM\ 26}$	62.48	86.82	92.41	98.87	85.15
ReCoVR (Our Interactive CoVR)					
Turn 0	54.30	77.00	84.35	95.77	77.86
Turn 1	**74.30**	87.28	93.08	98.75	**88.35**
Turn 2	80.32	**89.28**	**93.81**	98.90	90.45
Turn 3	84.39	91.28	94.56	99.18	92.35
Turn 4	87.72	93.04	95.27	**99.30**	93.83
Turn 5	90.22	94.13	96.24	99.57	95.04

Implementation Details. Unless specified otherwise, ReCoVR and the simulator use Qwen3-VL-4B (Bai et al., 2025a), and both retrieval channels use BLIP-ViT-Large (Li et al., 2022). We set the TW-RRF window W=5, RRF constant k=60, candidate pool K=100, and rank-cap positions κ_{neg}=11 and κ_{neu}=4. All models remain frozen; ReCoVR requires no training.

4.2 Main Comparison under the ICoVR Protocol

Table 1 compares interactive methods under the shared ICoVR protocol. ReCoVR leads in R@1 at every feedback turn, and its margin over the strongest baseline grows across turns. Paired session-level 95% confidence intervals for the per-turn R@1 advantage lie above zero (Appendix A.3).

The trajectory analysis in Appendix A.4 explains why the margin grows. UMIVR fails to improve the target rank in more than half of its active feedback turns, far more often than ReCoVR, and its failures tend to persist: after a failed turn, the rank improves on the next turn in only 30.12% of cases, below UMIVR's own average. ReCoVR recovers in 88.15% of cases, above its own average. An open-loop system that stalls tends to stay stalled, whereas ReCoVR registers the failure in its ranking history and corrects it. These rates are computed offline on each method's own trajectories; target ranks are never available to the system during search.

Fig. 3 shows the same pattern on Dense-WebVid-CoVR and FineCVR. Dense-WebVid-CoVR pairs the same videos with much longer modification texts, so the continued gains show that interaction helps even when the initial query is already detailed.

Table 2 compares against single-turn CoVR (Ventura et al., 2024a), ECDE (Thawakar et al., 2024), CoVR-2 (Ventura et al., 2024b), TFR-CVR (Hummel et al., 2024), FDCA (Yue et al., 2025), HUD (Chen et al., 2025b), CADP (Wang et al., 2025), and REFINE (Hu et al., 2026). Starting from a mid-range backbone, ReCoVR surpasses the best single-turn method in R@1 after a single feedback turn and keeps improving with further interaction.

4.3 Ablation Study

Component contributions. Table 3 adds components one at a time, and three trends emerge. First, the two channels are complementary: the CoVR-only and T2V-only variants perform similarly, yet routing feedback to both through the Intent Pathway improves over either alone, so each channel succeeds on feedback that the other handles poorly. Second, the two correction mechanisms add up. Rank-cap gives the larger single gain on both benchmarks, but without State Reflection σ_t defaults to neutral and rank-cap can only react to stagnation. Adding State Reflection supplies rejection signals and improves R@1 and BRI at every turn, and the gap to the Intent-only variant widens as the interaction proceeds. Rank-cap corrects the current ranking immediately, while State Reflection

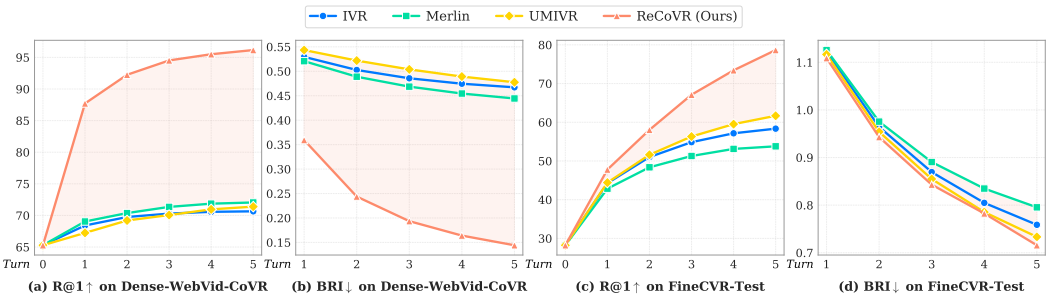

Figure 3: **ICoVR comparison on additional CoVR benchmarks.** Under the same interactive protocol, ReCoVR shows stronger turn-wise improvement than adapted I-TVR baselines on Dense-WebVid-CoVR-Test (a,b) and FineCVR-Test (c,d). The shaded region marks ReCoVR's margin over the baselines, which widens across turns in R@1.

Table 3: **Component ablation** on WebVid-CoVR-Test and FineCVR-Test. Each row enables the checked components with all other settings fixed; without State Reflection, the satisfaction signal σ_t is set to neutral. Dual-channel routing, State Reflection, and rank-cap each contribute, and the full model has the best R@1 at every turn on both benchmarks.

CoVR Channel	*T2V Channel*	*Intent Pathway*	*State Reflection*	*Rank-cap*	**Dataset**	**WebVid-CoVR-Test**					**FineCVR-Test**				
					Metrics	*Turn 1*	*2*	*3*	*4*	*5*	*Turn 1*	*2*	*3*	*4*	*5*
✓					R@1↑	68.15	71.64	72.73	73.40	74.52	44.54	51.92	56.28	59.60	61.34
					BRI↓	0.6938	0.5699	0.5192	0.4890	0.4679	1.1249	0.9576	0.8607	0.7895	0.7487
	✓				R@1↑	66.76	71.67	72.42	73.08	73.59	43.39	49.70	54.21	58.56	60.55
					BRI↓	0.7061	0.5724	0.5210	0.4906	0.4694	1.1865	0.9579	0.8569	0.7962	0.7507
✓	✓	✓			R@1↑	69.29	72.69	74.61	76.17	77.46	45.56	52.17	58.85	61.22	63.69
					BRI↓	0.6915	0.5619	0.5017	0.4733	0.4590	1.1151	0.9556	0.8518	0.7859	0.7310
✓			✓		R@1↑	68.94	73.63	75.36	76.28	78.60	45.42	52.71	57.94	62.23	65.38
					BRI↓	0.6935	0.5608	0.5114	0.4860	0.4533	1.1243	0.9556	0.8584	0.7897	0.7486
✓			✓	✓	R@1↑	71.61	78.11	81.46	83.65	85.17	46.85	56.13	62.34	69.97	73.89
					BRI↓	0.6804	0.5445	0.4832	0.4330	0.4019	1.1204	0.9553	0.8544	0.7872	0.7380
✓	✓	✓	✓		R@1↑	70.42	75.51	78.44	80.13	83.26	46.84	53.24	61.14	64.95	69.40
					BRI↓	0.6918	0.5491	0.4912	0.4691	0.4377	1.1143	0.9527	0.8486	0.7858	0.7284
✓	✓	✓		✓	R@1↑	72.41	79.39	83.94	86.70	87.52	47.48	55.66	64.17	71.85	76.87
					BRI↓	0.6810	0.5355	0.4847	0.4241	0.3964	1.1140	0.9531	0.8453	0.7841	0.7171
✓	✓	✓	✓	✓	R@1↑	74.30	80.32	84.39	87.72	90.22	47.78	58.04	67.11	73.42	78.65
					BRI↓	0.6737	0.5271	0.4538	0.4034	0.3643	1.1087	0.9529	0.8430	0.7825	0.7157

shapes later queries. Third, the gains do not come from the extra retriever: with the CoVR channel alone, State Reflection and rank-cap already surpass every adapted baseline at every feedback turn.

Simpler uses of history. A baseline that regenerates one composed query from the full interaction history, without reflection, improves over turns but falls increasingly behind ReCoVR, and its R@50 even drops below the Turn-0 level. Re-deriving a single query from a growing history thus tends to over-condition the ranking. Transplanting rank-cap into IVR and UMIVR helps both, yet their gap to ReCoVR still widens over turns (Appendix A.2). Appendix A.7 compares further single-mechanism strategies, all of which trail the full loop.

Fusion and anchoring. TW-RRF gives the best final retrieval among the fusion variants (Appendix A.1). Uniform turn weights come close early but fall behind in later turns, current-turn-only fusion is competitive only at Turn 1, and score-level fusion is weaker throughout because raw similarities are not comparable across channels. Cross-turn evidence and recency weighting therefore both matter. Dynamic anchoring outperforms a fixed reference at every turn, since each edit then refers to the video the user has just seen (Appendix A.6).

Sensitivity. Performance is stable across the tested k, K, and rank-cap positions (Fig. 4), and a larger window W helps until gains saturate, indicating that recent history is useful while much older turns add little. Rewording all module prompts barely changes final accuracy, and nearly all structured outputs parse on the first attempt (Appendix A.5).

Generalization. Appendix B shows gains on FashionIQ, across nine backbone configurations and two reasoning-VLM families, and on I-TVR (MSRVTT, AVSD) when reflection is added to UMIVR under its original protocol, so reflection also helps existing pipelines.

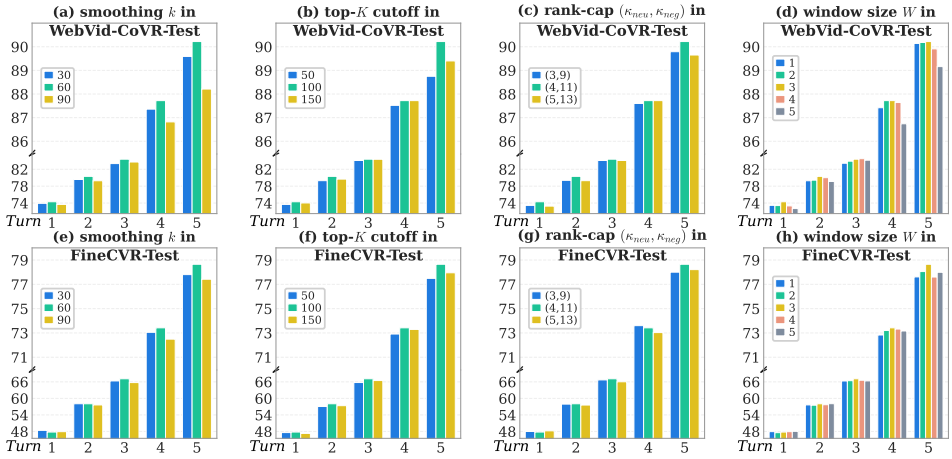

Figure 4: **Parameter sensitivity of TW-RRF** on WebVid-CoVR-Test (top) and FineCVR-Test (bottom). ReCoVR remains stable across broad ranges of k, top-K cutoff, rank-cap positions, and window size W, indicating that its gains do not depend on precise parameter tuning.

Table 4: **Human versus simulated feedback on WebVid-CoVR-Test.** On 100 ICoVR instances, ReCoVR follows a closely matching improvement trend under free-form human feedback and simulated feedback. IVR with simulated feedback is shown for reference.

Method	User Type	R@1↑				R@5↑				BRI↓			Feedback Words(Avg.)
		Turn 0	*1*	*2*	*3*	*0*	*1*	*2*	*3*	*1*	*2*	*3*	
IVR	Simulator	43.0	58.0	58.0	59.0	63.0	75.0	82.0	81.0	1.005	0.860	0.788	6.1
ReCoVR	Simulator	43.0	63.0	73.0	79.0	63.0	83.0	83.0	85.0	0.943	0.751	0.650	6.3
ReCoVR	Real Humans	43.0	62.0	70.0	76.0	63.0	77.0	80.0	84.0	0.977	0.800	0.700	13.7

4.4 Human Feedback and Computational Cost

Simulated feedback is reproducible (Zhang et al., 2024; Liang & Albanie, 2023; Levy et al., 2023; Zhang et al., 2025) but follows a fixed policy, so we check whether it reflects real interaction by running ReCoVR with free-form feedback from real users, following the protocol of Liang & Albanie (2023). Five participants interact with ReCoVR on 100 WebVid-CoVR-Test instances for up to three turns each, through a web interface that displays the target video, the current top-ranked result, and any system question (Appendix D.3). Table 4 shows closely matching trajectories under human and simulated feedback, both well above IVR with simulated feedback. Human feedback is roughly twice as long and less templated than simulated feedback, yet ReCoVR turns it into comparable gains (protocol in Appendix D). ReCoVR also retains its advantage under simulated feedback with noisy wording, omitted constraints, or modify-only edits (Appendix D.4). **Computational Cost.** On 100 WebVid-CoVR-Test instances with one RTX A6000 GPU (Appendix C), ReCoVR needs fewer feedback turns than UMIVR but more time per turn. VLM reasoning dominates this cost, while retrieval and rank fusion take only hundredths of a second per turn and the memory bank adds a few megabytes. Under a conservative estimate of human response time, ReCoVR's own computation is about 15% of end-to-end latency, so the user rather than the model sets the pace of practical interaction, and a smaller or faster reasoning model would reduce it further.

5 Conclusion

We introduced ICoVR and showed that adapted interactive retrievers often stall because they ignore their own ranking history. ReCoVR closes this loop without training and improves over these baselines on three CoVR benchmarks, with a margin that grows across turns. More broadly, a retriever's own ranking history is an inexpensive and underused signal: it needs no access to the target and no retraining, yet it identifies the turns at which feedback alone fails. Next steps include adaptive channel weighting for large reference–target gaps (Appendix E), lighter reasoning models for lower per-turn cost, and user studies with more diverse participants and search behaviors.

Reproducibility Statement

The experimental setup is described in Section 4.1. Appendix F provides the turn-level algorithm, Appendix G lists the prompts and output schemas, and Appendix D describes the user simulation and human-feedback protocols. Additional analyses and computational details are provided in Appendices A and C. We will release source code in the supplementary material to support reproduction.

Ethics Statement

The human-feedback study received institutional ethics approval. Participants were informed of the task procedure and minimal risks, and no demographic attributes or sensitive personal information were collected. Further details are provided in Appendix D.3. Video retrieval over sensitive or personal collections may raise privacy concerns; responsible deployment should consider data consent, access control, and dataset and model biases, as discussed in Appendix H.

AI Use Statement

LLMs assisted with manuscript drafting and language polishing, as well as the implementation and checking of parts of the code. The authors reviewed the AI-assisted work and remain responsible for the accuracy and integrity of the manuscript and code.

References

Lorenzo Agnolucci, Alberto Baldrati, Alberto Del Bimbo, and Marco Bertini. isearle: Improving textual inversion for zero-shot composed image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7558–7567, 2019.

Mohammad Aliannejadi, Jacek Gwizdka, and Hamed Zamani. Interactions with generative information retrieval systems. In *Information Access in the Era of Generative AI*, pp. 47–71. Springer, 2024.

Leif Azzopardi, Maarten De Rijke, and Krisztian Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 455–462, 2007.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3952–3962, 2025b.

Krisztian Balog and ChengXiang Zhai. User simulation for evaluating information access systems. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 302–305, 2023.

Keping Bi, Qingyao Ai, Yongfeng Zhang, and W Bruce Croft. Conversational product search based on negative feedback. In *Proceedings of the 28th acm international conference on information and knowledge management*, pp. 359–368, 2019.

Guanyu Cai, Jun Zhang, Xinyang Jiang, Yifei Gong, Lianghua He, Fufu Yu, Pai Peng, Xiaowei Guo, Feiyue Huang, and Xing Sun. Ask&confirm: active detail enriching for cross-modal retrieval with partial query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1835–1844, 2021.

Ke Chen, Shengyuan Han, Yongfeng Huang, Yujin Zhu, Jingwei Xiong, Liang Xu, and Jundong Liu. Adept: An entropy-driven dual-strategy agent for interactive video retrieval. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9442–9446. IEEE, 2026.

Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *European Conference on Computer Vision*, pp. 136–152. Springer, 2020.

Yanzhe Chen, Zhiwen Yang, Jinglin Xu, and Yuxin Peng. Mai: A multi-turn aggregation-iteration model for composed image retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Haokun Wen, and Weili Guan. Hud: Hierarchical uncertainty-aware disambiguation network for composed video retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6143–6152, 2025b.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 758–759, 2009.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. Trec cast 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624*, 2020.

Chadha Degachi, Ujjayan Dhar, Evangelos Niforatos, and Gerd Kortuem. Towards a domain expert evaluation framework for conversational search in healthcare. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–9, 2025.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. Simulatorarena: Are user simulators reliable proxies for multi-turn evaluation of ai assistants? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35200–35278, 2025.

Pierre Erbacher, Ludovic Denoyer, and Laure Soulier. Interactive query clarification and refinement via user simulation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2420–2425, 2022.

Yanglin Feng, Yongxiang Li, Yuan Sun, Yang Qin, Dezhong Peng, and Peng Hu. Interactive cross-modal learning for text-3d scene retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31, 2018.

Animesh Gupta, Jay Parmar, Ishan Rajendrakumar Dave, and Mubarak Shah. From play to replay: Composed video retrieval for temporally fine-grained videos. *arXiv preprint arXiv:2506.05274*, 2025.

Donghoon Han, Eunhwan Park, Gisang Lee, Adam Lee, and Nojun Kwak. Merlin: Multimodal embedding refinement via llm-based iterative navigation for text-video retrieval-rerank pipeline. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 547–562, 2024.

Ning Han, Yawen Zeng, Shaohua Long, Chengqing Li, Sijie Yang, Dun Tan, Jianfeng Dong, and Jingjing Chen. Ivcr-200k: A large-scale multi-turn dialogue benchmark for interactive video corpus retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 127–137, 2025.

Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pp. 1463–1471, 2017.

Chen He, Shenshen Li, Zheng Wang, Hua Chen, Fumin Shen, and Xing Xu. Chatting with interactive memory for text-based person retrieval. *Multimedia Systems*, 31(1):31, 2025.

Yupeng Hu, Zixu Li, Zhiwei Chen, Qinlei Huang, Zhiheng Fu, Mingzhu Xu, and Liqiang Nie. Refine: Composed video retrieval via shared and differential semantics enhancement. *ACM Trans. Multimedia Comput. Commun. Appl.*, February 2026. ISSN 1551-6857. doi: 10.1145/3796712. Just Accepted.

Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Junyang Ji, Shengjun Zhang, Da Li, Yuxiao Luo, Yan Wang, Di Xu, Biao Yang, Wei Yuan, Fan Yang, Zhihai He, and Wenming Yang. OmniCVR: A benchmark for omni-composed video retrieval with vision, audio, and text. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KxxR7emO5K.

Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval*, 3(1–2):1–224, 2009.

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. 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)*, pp. 791–809, 2024.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv*, 2026.

Kaiqu Liang and Samuel Albanie. Simple baselines for interactive video retrieval with questions and answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11091–11101, 2023.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2125–2134, 2021.

Yiding Lu, Mouxing Yang, Dezhong Peng, Peng Hu, Yijie Lin, and Xi Peng. Llava-reid: Selective multi-image questioner for interactive person re-identification. *arXiv preprint arXiv:2504.10174*, 2025.

Pengfei Luo, Jingbo Zhou, Tong Xu, Yuan Xia, Linli Xu, and Enhong Chen. Imagescope: Unifying language-guided image retrieval via large multimodal model collective reasoning. In *Proceedings of the ACM on Web Conference 2025*, pp. 1666–1682, 2025.

Chenyang Lyu, Manh-Duy Nguyen, Van-Tu Ninh, Liting Zhou, Cathal Gurrin, and Jennifer Foster. Dialogue-to-video retrieval. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (eds.), *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part II*, volume 13981 of *Lecture Notes in Computer Science*, pp. 493–501. Springer, 2023. doi: 10.1007/978-3-031-28238-6_40.

Avinash Madasu, Junier Oliva, and Gedas Bertasius. Learning to retrieve videos by asking questions. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (eds.), *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pp. 356–365. ACM, 2022. doi: 10.1145/3503161.3548361.

Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. A survey of conversational search. *ACM Transactions on Information Systems*, 43(6):1–50, 2025.

Anwesan Pal, Sahil Wadhwa, Ayush Jaiswal, Xu Zhang, Yue Wu, Rakesh Chada, Pradeep Natarajan, and Henrik I Christensen. Fashionntm: Multi-turn fashion image retrieval via cascaded memory. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11323–11334, 2023.

Jenny Preece, Yvonne Rogers, Helen Sharp, David Benyon, Simon Holland, and Tom Carey. *Human-computer interaction*. Addison-Wesley Longman Ltd., 1994.

Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, and Peng Hu. Human-centered interactive learning via mllms for text-to-image person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14390–14399, 2025.

Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Benchmarking robustness of multimodal image-text models under distribution shift. *arXiv preprint arXiv:2212.08044*, 2022.

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*, pp. 8748–8763. PMLR, 2021.

Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on circuits and systems for video technology*, 8(5):644–655, 1998.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Omkar Thawakar, Muzammal Naseer, Rao Muhammad Anwer, Salman Khan, Michael Felsberg, Mubarak Shah, and Fahad Shahbaz Khan. Composed video retrieval via enriched context and discriminative embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26896–26906, 2024.

Omkar Thawakar, Dmitry Demidov, Ritesh Thawkar, Rao Muhammad Anwer, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. Beyond simple edits: Composed video retrieval with dense modifications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20435–20444, 2025.

Omkar Thawakar, Dmitry Demidov, Vaishnav Potlapalli, Sai Prasanna Teja Reddy Bogireddy, Viswanatha Reddy Gajjala, Alaa Mostafa Lasheen, Rao Muhammad Anwer, and Fahad Khan. Covr-r: Reason-aware composed video retrieval. *arXiv preprint arXiv:2603.20190*, 2026.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5270–5279, 2024a.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr-2: Automatic data construction for composed video retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11409–11421, 2024b.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6439–6448, 2019.

Hao Wang, Fang Liu, Licheng Jiao, Jiahao Wang, Shuo Li, Lingling Li, Puhua Chen, and Xu Liu. Vision-by-prompt: Context-aware dual prompts for composed video retrieval. *Pattern Recognition*, pp. 112378, 2025.

Xuanhui Wang, Hui Fang, and ChengXiang Zhai. A study of methods for negative relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 219–226, 2008.

Zhenduo Wang, Zhichao Xu, Vivek Srikumar, and Qingyao Ai. An in-depth investigation of user response simulation for conversational search. In *Proceedings of the ACM Web Conference 2024*, pp. 1407–1418, 2024.

Hao Wei, Shuhui Wang, Zhe Xue, Shengbo Chen, and Qingming Huang. Conversational composed retrieval with iterative sequence refinement. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 6390–6399, 2023.

Hong Wu, Hanqing Lu, and Songde Ma. Willhunter: interactive image retrieval with multilevel relevance. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pp. 1009–1012. IEEE, 2004.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11307–11317, 2021.

Jiaxin Wu, Chong-Wah Ngo, Wing-Kwong Chan, Sheng-Hua Zhong, Xiong-Yong Wei, and Qing Li. Multimodal llm-based query paraphrasing for video search. *arXiv preprint arXiv:2407.12341*, 2024.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 5288–5296. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.571.

Zhongyu Yang, Wei Pang, and Yingfang Yuan. Xr: Cross-modal agents for composed image retrieval. *arXiv preprint arXiv:2601.14245*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 471–487, 2018.

Yifei Yuan and Wai Lam. Conversational fashion image retrieval via multiturn natural language feedback. In *Proceedings of the 44th International ACM SIGIR conference on research and development in information retrieval*, pp. 839–848, 2021.

WU Yue, Zhaobo Qi, Yiling Wu, Junshu Sun, Yaowei Wang, and Shuhui Wang. Learning fine-grained representations through textual token disentanglement in composed video retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bingqing Zhang, Zhuo Cao, Heming Du, Yang Li, Xue Li, Jiajun Liu, and Sen Wang. Quantifying and narrowing the unknown: Interactive text-to-video retrieval via uncertainty minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22120–22130, 2025.

Erhan Zhang, Xingzhu Wang, Peiyuan Gong, Yankai Lin, and Jiaxin Mao. Usimagent: Large language models for simulating search users. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2687–2692, 2024.

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

A Additional Analyses and Ablations . . . 16
A.1 Rank Fusion Strategy . . . 16
A.2 Baseline Rank-cap Transfer . . . 17
A.3 Statistical Significance via Paired Bootstrap . . . 17
A.4 Measuring Open-Loop Failure Modes . . . 18
A.5 Robustness to Prompt Rewording and Structured-Output Reliability . . . 18
A.6 Fixed-Reference versus Dynamic Anchoring . . . 19
A.7 Simplified Multi-Turn Strategies . . . 20
B Generalization . . . 21
C Efficiency Analysis . . . 22
D User Simulation and Human-Feedback Study Details . . . 24
D.1 Motivation: Why Simulate Users? . . . 24
D.2 User Simulator Design . . . 25
D.3 Human-Feedback Interface . . . 25
D.4 Robustness to Degraded Simulated Feedback . . . 26
E Qualitative Case Studies . . . 28
F Algorithmic Overview of ReCoVR . . . 28
G Prompt Templates . . . 29
H Limitations and Discussion . . . 31

A Additional Analyses and Ablations

A.1 Rank Fusion Strategy

The main paper adopts Time-Weighted Reciprocal Rank Fusion (TW-RRF) to aggregate candidate lists across channels and turns. Here we compare it against five alternatives on WebVid-CoVR-Test to justify this choice: (1) Static-RRF fuses only the current turn's two channel lists using reciprocal rank with fixed channel weights, discarding all historical evidence; (2) Uniform-RRF aggregates across a sliding window like TW-RRF but assigns equal weight to every turn; (3) Penalty-RRF replaces the additive indicator with a penalty rank r_{pen} for absent candidates, so that missing from a list incurs a small negative contribution rather than zero; (4) SimMax performs score-level fusion by taking the maximum raw similarity across channels and turns; (5) SimSum sums the raw similarity scores instead.

As shown in Table 5, TW-RRF achieves the best R@1 from Turn 2 onward and the strongest final performance on every metric. Static-RRF is marginally ahead at Turn 1 but plateaus quickly because it cannot accumulate cross-turn evidence. Uniform-RRF performs close to TW-RRF in early turns, with slightly lower BRI at Turns 2–3, but falls behind as the interaction progresses, indicating that recency weighting matters more once later turns carry more refined intent. Penalty-RRF underperforms both Uniform-RRF and TW-RRF, indicating that penalizing absent candidates is overly conservative and suppresses useful partial evidence from individual channels. The two score-level baselines (SimMax, SimSum) are consistently weaker, which is expected because raw similarity scales differ across channels and turns, making direct score comparison unreliable. These results validate the three design choices behind TW-RRF: multi-turn aggregation, recency weighting, and rank-level (rather than score-level) fusion.

Table 5: **Ablation on rank fusion strategies** on WebVid-CoVR-Test. We compare TW-RRF against three rank-level variants (Static-RRF, Uniform-RRF, Penalty-RRF) and two score-level baselines (SimMax, SimSum). TW-RRF has the best R@1 from Turn 2 onward and the best Turn-5 result on every metric; Static-RRF is marginally ahead in R@1 at Turn 1, and Uniform-RRF has slightly lower BRI at Turns 2–3.

Rank Fusion Strategy	R@1↑					R@5↑					BRI↓				
	Turn 1	*2*	*3*	*4*	*5*	*Turn 1*	*2*	*3*	*4*	*5*	*Turn 1*	*2*	*3*	*4*	*5*
Penalty-RRF	70.11	75.77	80.13	84.59	87.79	86.66	88.54	89.63	90.88	92.94	0.6944	0.5614	0.4885	0.4321	0.3877
Static-RRF	74.36	78.33	80.95	83.76	86.03	87.25	86.93	87.56	88.81	90.10	0.6701	0.5557	0.4915	0.4490	0.4164
Uniform-RRF	73.63	79.81	83.57	86.79	89.03	88.38	88.71	90.69	91.98	93.19	0.6726	0.5250	0.4512	0.4106	0.3736
SimMax	68.58	69.76	75.63	80.16	82.75	85.52	84.47	85.84	87.28	88.58	0.7162	0.6052	0.5520	0.5092	0.4738
SimSum	69.64	73.32	75.35	76.96	83.29	87.56	89.98	90.65	91.28	92.61	0.6938	0.5583	0.4933	0.4526	0.4181
TW-RRF	74.30	80.32	84.39	87.72	90.22	87.28	89.28	91.28	93.04	94.13	0.6737	0.5271	0.4538	0.4034	0.3643

Table 6: **Effect of transplanting Rank-cap into open-loop interactive baselines** on WebVid-CoVR-Test. Turn 0 is omitted because all methods share the same initialization. While Rank-cap consistently improves both `IVR` and `UMIVR`, ReCoVR remains the best across all turns, indicating that rank-space intervention alone does not account for the full gains.

	Methods	*Turn 1*	*2*	*3*	*4*	*5*
R@1↑	IVR	66.55	70.11	71.95	73.00	73.51
	IVR + Rank-cap	73.04	76.64	79.89	80.75	82.08
	UMIVR	67.10	71.09	72.85	74.10	75.04
	UMIVR + Rank-cap	72.81	77.03	80.63	81.77	83.18
	ReCoVR (ours)	**74.30**	**80.32**	**84.39**	**87.72**	**90.22**
BRI↓	IVR	0.7368	0.6187	0.5560	0.5161	0.4880
	IVR + Rank-cap	0.7074	0.5715	0.4985	0.4516	0.4187
	UMIVR	0.7285	0.6049	0.5394	0.4986	0.4700
	UMIVR + Rank-cap	0.6967	0.5555	0.4798	0.4314	0.3982
	ReCoVR (ours)	**0.6737**	**0.5271**	**0.4538**	**0.4034**	**0.3643**

A.2 Baseline Rank-cap Transfer

To isolate the contribution of the Rank-cap policy, we transplant the same rule into the open-loop interactive baselines `IVR` and `UMIVR`. Their original single-channel query-update pipelines remain unchanged. We reuse the same cap positions as ReCoVR (κ_{neg}=11, κ_{neu}=4) and derive the satisfaction signal from the simulator's action type, which gives the baselines a rejection signal that ReCoVR must instead infer from the feedback, applying demotion as a lightweight post-processing step without introducing the Intent Pathway or State Reflection.

As shown in Table 6, Rank-cap yields consistent gains for both baselines across all turns: at Turn 5, R@1 improves by 8.57 points for `IVR` and 8.14 points for `UMIVR`, with corresponding BRI improvements. This confirms that rank-space intervention is a useful correction mechanism that transfers readily to existing interactive pipelines. ReCoVR achieves the best performance at every turn, and the margin grows over successive rounds: the gap over UMIVR + Rank-cap is 1.49 points at Turn 1 and 7.04 points at Turn 5. This widening trend is consistent with the component ablation in Table 3: Rank-cap provides effective immediate intervention, while the dual-channel Intent Pathway and trajectory-aware State Reflection contribute cumulative gains that become increasingly important as interaction history grows. The full architecture combines these complementary mechanisms to sustain improvement across turns.

A.3 Statistical Significance via Paired Bootstrap

The main results in Table 1 are point estimates. To attach uncertainty to the comparison against the strongest adapted baseline, we rerun ReCoVR and UMIVR in a single matched environment and pair all 2,556 test-session trajectories. Both methods start from the identical Turn-0 ranking, so every session yields a paired difference. We then resample complete session trajectories 10,000 times. Each resample uses the same session indices for both methods. We report percentile 95% intervals on the per-turn difference in R@1.

Table 7 reports the outcome. The paired improvement grows from $+6.77$ points at Turn 1 to $+14.32$ points at Turn 5, and every interval lies entirely above zero. This matched rerun reproduces the ordering and turn-wise trend of Table 1 to within 0.55 points for both methods, so we retain the original numbers as the main result and use the rerun only to quantify uncertainty. Two scope notes apply. The intervals characterize test-session sampling variability conditional on a fixed retriever, reasoning model, simulator, and interaction protocol; they do not capture variability from re-drawing model or simulator behavior. And the five turns are successive states of the same sessions rather than independent samples, so we report per-turn intervals instead of a single aggregate test.

Table 7: **Paired session-level uncertainty** for ReCoVR over UMIVR on WebVid-CoVR-Test, from a matched rerun of both methods over 2,556 paired session trajectories with 10,000 bootstrap resamples. The improvement widens across turns and every 95% interval lies entirely above zero.

	Turn 1	*Turn 2*	*Turn 3*	*Turn 4*	*Turn 5*
UMIVR **R@1** ↑	67.21	71.56	73.40	74.61	75.35
ReCoVR **R@1** ↑	73.98	79.93	85.05	88.18	89.67
ΔR@1 (pp)	**+6.77**	**+8.37**	**+11.66**	**+13.58**	**+14.32**
Paired 95% CI	[+5.20, +8.29]	[+6.77, +9.90]	[+10.09, +13.22]	[+12.01, +15.14]	[+12.75, +15.88]

A.4 Measuring Open-Loop Failure Modes

Our motivation rests on an empirical premise: open-loop interactive retrieval drifts, stagnates, and keeps returning candidates the user has already rejected. We measure these behaviors directly rather than inferring them from end-task metrics. We use the matched rerun of Appendix A.3 to score every *active* feedback transition. A transition is active when the simulated user is still interacting and has not issued a stop action. We compare the target rank before and after each active turn. Turn 0, preparation steps, and carry-forward rows after an early stop are excluded. Because ReCoVR terminates more sessions earlier, the two methods enter interaction on the same 1,167 sessions but execute a different number of active transitions (3,010 versus 4,061), so all rates are computed against each method's own denominator.

Table 8 reports the outcome, with each failure mode operationalized as a measurable proxy on the target rank. We define query drift as a turn in which the target rank strictly worsens. It occurs in 18.52% of UMIVR's active transitions versus 12.69% of ReCoVR's. Stagnation means that feedback fails to improve the target rank. Its rate is 54.25% for UMIVR versus 20.23% for ReCoVR. A rejected candidate is repeated when the top-1 result remains unchanged after non-stop feedback. This occurs in 70.77% of UMIVR's transitions. ReCoVR has no rate to report for this failure mode because rank-cap demotes the rejected top-1 at every such turn and rules it out by design.

The more informative question is what happens *after* a bad turn. Restricting to cases where the session continues and another feedback turn is executed, ReCoVR improves the target rank on the following turn in 88.15% of cases against 30.12% for UMIVR. This contrast should be read against each method's own base rate: ReCoVR improves the rank in 79.77% of its turns overall and UMIVR in 45.75%, so recovering at 88.15% puts ReCoVR 8.38 points *above* its own average, while 30.12% puts UMIVR 15.63 points *below* its own. This reversal in sign distinguishes a closed loop from an open one more clearly than the raw ratio: a stalled open-loop system tends to stay stalled, because nothing in its update rule registers that the previous turn failed. Finally, the rank-cap policy fires in 49.20% of ReCoVR's active transitions, and 76.57% of those transitions show a same-turn improvement in target rank. The same turn also performs intent decomposition, retrieval, and fusion, so this is an association attached to the triggering condition rather than an isolated causal effect of rank-cap; the matched ablations of Sec. 4.3 carry the component-level attribution.

A.5 Robustness to Prompt Rewording and Structured-Output Reliability

Intent decomposition, query reformulation, state reflection, and question asking are all realized as structured prompts, so we test whether the results depend on a particular wording. We create two additional semantically equivalent prompt suites called Variant A and Variant B. Each rewrites all four module prompts. We rerun the full protocol on WebVid-CoVR-Test with the retriever, simulator, decoding settings, and interaction protocol held fixed.

Table 8: **Direct measurement of the open-loop failure modes** over active feedback transitions on WebVid-CoVR-Test. Rates use each method's own number of active transitions (ReCoVR 3,010, UMIVR 4,061). Each failure mode from Sec. 1 is operationalized as a measurable proxy. Repeated rejected candidates cannot occur under ReCoVR, whose rank-cap rule demotes a rejected top-1 after every non-stop feedback. The lower block conditions on transitions that failed to improve the rank and are followed by another executed turn.

Measured over active feedback transitions	UMIVR	ReCoVR
Query drift: the target rank strictly worsens ↓	18.52% (752/4,061)	**12.69% (382/3,010)**
Stagnation: the target rank does not improve ↓	54.25% (2,203/4,061)	**20.23% (609/3,010)**
Repeated rejected candidates: top-1 unchanged after feedback ↓	70.77%	*prevented by rank-cap*
After a turn that failed to improve the rank		
Target rank improves on the next turn ↑	30.12% (1,720 eligible)	**88.15% (557 eligible)**
relative to the method's own per-turn improvement rate	−15.63 pp	+8.38 **pp**
Median change in target rank on the next turn ↓	0	−2

Table 9 reports the three trajectories. All suites preserve monotonic multi-turn improvement, and Turn-5 R@1 spans only 0.39 points (89.95–90.34). The effect of rewording is concentrated in the early turns: Variant B loses 1.60 points at Turn 1 and then recovers to within 0.27 points by Turn 5. We therefore claim robustness of the final conclusion under semantically equivalent wording, not insensitivity to prompt design in general.

We separately quantify how often the structured-output interface itself fails. Across the three suites, 22,408 schema-constrained generations are issued, of which 22,304 are valid on the first pass (99.54%); the remaining 104 are recovered by a deterministic fallback, and no run terminates early or produces an unusable session. The Intent Decomposer and State Reflector drive the interaction loop. They produce valid outputs on the first pass in 18,903 of 18,904 calls (99.995%). Reliability is not uniform across modules: under Variant B the T2V initializer falls to 91.18% first-pass validity, which is precisely the case the fallback exists to absorb. These statistics characterize output *interface* reliability by measuring whether a decision can be parsed. They do not measure the semantic correctness of each decision. End-task metrics assess the resulting retrieval performance.

Table 9: **Robustness to prompt rewording** on WebVid-CoVR-Test. Variants A and B reword all four reasoning-module prompts while preserving their meaning. R@1 is reported per turn; the Turn-5 spread across the three suites is 0.39 points.

Prompt suite	*Turn 1*	*2*	*3*	*4*	*5*	*Turn-5* BRI↓
Original (main paper)	74.30	80.32	84.39	87.72	90.22	0.3643
Variant A	74.26	79.73	84.12	87.37	90.34	0.3697
Variant B	72.70	78.76	83.02	86.90	89.95	0.3846

A.6 Fixed-Reference versus Dynamic Anchoring

The CoVR channel anchors each turn on the previously fused top-1 result $\hat{L}_{t-1}(1)$ rather than on the original reference video v_{ref}. The motivation is that a `Modify` instruction describes how the *currently displayed* result should change, so applying it to the original reference would attach the edit to the wrong referent. We verify this choice with a matched comparison in which only the anchor differs: the simulator, reasoning VLM, retriever, both pathways, rank-cap, TW-RRF, turn budget, and stopping rule are unchanged, and each setting receives feedback generated online against its own displayed result rather than a replay of the other's dialogue.

Table 10 shows that dynamic anchoring is better at all five interactive turns, by $+1.52$ to $+4.10$ points of R@1 and $+2.99$ on average. The advantage generally grows in later turns. It peaks at Turn 4 and narrows slightly at Turn 5. Both settings improve monotonically over turns, so a fixed reference does not break interactive refinement; it simply converges lower. We do not report an oracle-anchor variant, since selecting the anchor from target-side information is not available at retrieval time and would serve only as an upper bound.

Table 10: **Anchor choice for the CoVR channel** on WebVid-CoVR-Test (R@1). Only the visual anchor differs between the two settings; each receives online feedback on its own displayed result.

Anchor for the CoVR channel	*Turn 1*	*2*	*3*	*4*	*5*
Fixed reference v_{ref}	72.78	77.93	81.25	83.62	86.41
Dynamic: previous fused top-1 (ours)	**74.30**	**80.32**	**84.39**	**87.72**	**90.22**
Δ	+1.52	+2.39	+3.14	+4.10	+3.81

A.7 Simplified Multi-Turn Strategies

Sec. 4.3 compares ReCoVR against a baseline that rewrites the complete interaction history into a single composed query at every turn. Table 11 places that baseline alongside the other simplified strategies studied in this appendix, each of which keeps one mechanism of the closed loop and discards the rest. There is little history to exploit at Turn 1, and all variants are competitive at this stage. They fall behind ReCoVR as interaction proceeds. Current-turn-only fusion is marginally ahead at Turn 1 but falls 4.19 points behind by Turn 5. Table 12 then gives the full trajectory of the history-rewriting baseline.

Table 11: **Simplified multi-turn strategies** on WebVid-CoVR-Test. Each row retains one closed-loop mechanism and discards the rest. The retained mechanisms are history aggregation, rank-space correction, and single-turn fusion.

Strategy	*Turn 1*	*Turn 5*	Gap at Turn 5
Full-history query rewriting (CoVR-only, no Reflection)	65.54	77.43	-12.79
Rank-cap transplanted onto UMIVR (Appendix A.2)	72.81	83.18	-7.04
ReCoVR without Rank-cap (Sec. 4.3)	70.42	83.26	-6.96
Current-turn-only rank fusion, no cross-turn evidence (Appendix A.1)	74.36	86.03	-4.19
ReCoVR	**74.30**	**90.22**	—

The baseline uses the same retriever, reasoning VLM, simulator, and interaction protocol, retrieves through the CoVR channel only, and has no Reflection Pathway; it differs from the rolling CoVR-only variant of Table 3, which updates the previous rewritten query incrementally rather than regenerating it from the whole history. Its R@1 improves across turns, so history does carry usable signal, but its gains at deeper cutoffs are small: R@5 stays below ReCoVR's Turn-1 value at every turn, R@10 stays nearly flat, R@50 falls below its Turn-0 value, and BRI improves only slowly. Rederiving a single query from an ever-longer history tends to over-condition the ranking on whichever constraints the model chooses to surface, which is the failure that trajectory-level diagnosis and rank-level fusion are designed to avoid.

Table 12: **Full-history rewriting baseline** on WebVid-CoVR-Test: same setup as ReCoVR, CoVR channel only, no Reflection Pathway, with the query regenerated from the complete interaction history at every turn. Deltas are against ReCoVR under the matched run.

Metric	*Turn 0*	*1*	*2*	*3*	*4*	*5*
R@1 $\uparrow$	54.30	65.54	70.58	73.24	75.04	77.43
R@5 $\uparrow$	77.00	76.56	78.25	78.96	80.95	82.94
R@10 $\uparrow$	84.35	83.22	84.15	83.10	84.94	85.56
R@50 $\uparrow$	95.77	95.19	94.13	94.25	93.63	93.04
BRI $\downarrow$	–	0.7816	0.6806	0.6160	0.5689	0.5309
ΔR@1 vs. ReCoVR	0.00	-8.76	-9.74	-11.15	-12.68	-12.79

B GENERALIZATION

Because ReCoVR is training-free and operates above frozen retrievers, its usefulness depends on whether the gains survive changes of domain, retrieval backbone, task, and reasoning model; we evaluate each in turn.

Generalization across domains. Table 13 evaluates cross-domain transfer on FashionIQ (Wu et al., 2021), a composed image retrieval benchmark with three fashion categories. Although ReCoVR starts from a weaker Turn 0 retriever than recent single-turn CIR methods, interactive feedback steadily improves performance and surpasses them after several turns. Compared with the adapted interactive baseline UMIVR, ReCoVR achieves larger and more sustained gains across all categories, while UMIVR tends to plateau after early turns. This suggests that closed-loop correction is not tied to video-specific cues, but can also support fine-grained composed image retrieval.

Table 13: **Generalization to composed image retrieval** on FashionIQ. ReCoVR consistently outperforms single-turn or interactive methods across all categories.

Methods		Dress				Shirts				Tops&Tees			
		R@1↑	**R@10↑**	**R@50↑**	**BRI↓**	**R@1↑**	**R@10↑**	**R@50↑**	**BRI↓**	**R@1↑**	**R@10↑**	**R@50↑**	**BRI↓**
Single-Turn Composed Image Retrieval													
CoVR-2 (Ventura et al., 2024b)$_{TPAMI\ 24}$		-	46.53	69.60	-	-	51.23	70.64	-	-	52.14	73.27	-
iSHARLE (Agnolucci et al., 2025)$_{TPAMI\ 25}$		-	31.81	50.20	-	-	24.19	45.12	-	-	31.72	53.29	-
X^R (Yang et al., 2026)$_{WWW\ 26}$		-	38.91	56.82	-	-	28.71	52.50	-	-	43.91	62.57	-
REFINE (Hu et al., 2026)$_{TOMM\ 26}$		-	50.32	72.98	-	-	57.90	77.87	-	-	59.82	80.83	-
Interactive Refinement (BLIP-ViT-Large)													
UMIVR	*Turn 0*	8.73	26.18	45.46	-	9.03	30.32	47.25	-	11.22	31.36	52.83	-
	Turn 1	15.32	38.47	60.59	3.535	16.39	41.51	60.94	3.530	18.05	44.01	64.71	3.290
	Turn 2	18.59	40.80	62.92	3.228	18.55	42.89	62.81	3.229	20.40	45.49	66.96	2.998
	Turn 3	20.92	41.45	63.46	3.045	20.07	42.93	62.90	3.063	22.44	45.84	67.77	2.831
	Turn 4	22.76	40.70	62.92	2.926	20.76	42.69	63.35	2.954	23.41	46.46	68.28	2.717
	Turn 5	24.00	42.64	64.55	2.830	21.59	43.47	63.84	2.871	24.48	47.83	68.79	2.631
ReCoVR(Ours)	*Turn 0*	8.73	26.18	45.46	-	9.03	30.32	47.25	-	11.22	31.36	52.83	-
	Turn 1	23.30	49.23	71.54	3.257	26.74	49.56	71.15	3.268	26.77	53.54	74.55	3.035
	Turn 2	29.25	51.81	73.83	2.815	31.50	51.13	72.77	2.834	33.86	55.12	76.64	2.616
	Turn 3	35.75	56.52	77.14	2.580	37.54	56.82	76.79	2.603	40.13	60.17	80.01	2.393
	Turn 4	41.99	60.14	79.67	2.407	43.72	59.72	79.20	2.434	46.20	63.13	81.95	2.230
	Turn 5	**47.89**	**62.62**	**81.41**	**2.268**	**47.60**	**61.92**	**80.57**	**2.301**	**50.99**	**66.34**	**82.66**	**2.101**

Generalization across retrieval backbones. We examine whether ReCoVR depends on a particular retrieval encoder. Table 14 evaluates nine CoVR/T2V channel configurations on WebVid-CoVR-Test, including seven homogeneous settings where both channels use the same encoder and two heterogeneous settings where the channels use different encoders. These configurations cover CLIP (Radford et al., 2021), BLIP (Li et al., 2022), LanguageBind (Zhu et al., 2023), and Qwen3-VL-Embedding (Li et al., 2026) families.

All configurations show consistent improvement over turns, indicating that ReCoVR's gains are not tied to a specific retrieval backbone. Stronger encoders provide better starting points and higher final accuracy, while weaker encoders also benefit substantially from feedback, suggesting that closed-loop correction can compensate for limited initial retrieval capacity. Importantly, even when the Turn 0 retriever is already competitive with recent single-turn CoVR methods, ReCoVR still provides substantial additional gains, showing that **interactive refinement is complementary to advances in single-turn retrieval**. The heterogeneous settings further show that the dual-channel design does not require matched encoders and can also operate with encoder diversity.

Table 14: **ReCoVR with different retrieval backbones** on WebVid-CoVR-Test. We evaluate seven homogeneous configurations (same encoder for both channels) spanning five model families, plus two heterogeneous configurations (different encoders across channels). All configurations show consistent multi-turn improvement, confirming that ReCoVR is retrieval-backbone-agnostic.

CoVR Channel	T2V Channel	**R@1↑**					**R@5↑**					**BRI↓**				
		Turn 0	*1*	*2*	*3*	*4*	*5*	*Turn 0*	*1*	*2*	*3*	*4*	*5*	*Turn 1*	*2*	*3*	*4*	*5*
CLIP-Base	CLIP-Base	46.32	65.30	72.77	77.31	80.87	83.88	68.47	81.30	84.12	86.50	88.07	89.12	0.9293	0.7433	0.6791	0.5859	0.5379
CLIP-Large	CLIP-Large	55.36	73.67	80.91	84.90	88.18	90.02	77.78	87.40	90.06	91.90	93.35	94.21	0.6328	0.4961	0.4246	0.3761	0.3394
BLIP-Base	BLIP-Base	51.17	70.62	77.07	81.96	85.52	88.26	74.02	85.76	87.36	89.83	91.63	92.57	0.7598	0.6051	0.5263	0.4717	0.4291
BLIP-Large	BLIP-Large	54.30	74.30	80.32	84.39	87.72	90.22	77.00	87.28	89.28	91.28	93.04	94.13	0.6737	0.5271	0.4538	0.4034	0.3643
LanguageBind	LanguageBind	43.11	71.36	77.35	81.18	83.41	85.99	66.28	87.95	87.90	88.93	90.30	91.08	0.8897	0.6584	0.5585	0.4962	0.4504
Qwen3VL-Embed2B	Qwen3VL-Embed2B	59.47	79.97	84.00	88.34	91.20	92.72	81.30	92.14	93.35	94.64	95.74	96.60	0.5194	0.3965	0.3353	0.2931	0.2615
Qwen3VL-Embed8B	Qwen3VL-Embed8B	64.36	82.47	86.93	90.02	92.80	94.33	85.02	93.86	94.09	95.74	96.64	97.38	0.4262	0.3225	0.2711	0.2358	0.2091
BLIP-Large	LanguageBind	54.30	76.41	82.55	87.36	90.30	92.33	77.00	91.08	91.86	93.27	94.68	95.81	0.6408	0.4780	0.3988	0.3461	0.3071
Qwen3VL-Embed2B	BLIP-Base	59.47	75.74	82.75	86.62	89.32	91.74	81.30	89.16	91.71	93.43	94.76	95.70	0.5469	0.4317	0.3691	0.3256	0.2919

Table 15: **Generalization to I-TVR** on MSRVTT-1kA and AVSD-1k-Test. Adding the Reflection Pathway to UMIVR improves retrieval and outperforms all baselines on both video benchmarks.

Methods		*Turn 0*	*1*	*2*	*3*	*4*	*5*
MSRVTT-1kA Dataset							
R@1↑	IVR	43.1	61.4	65.8	68.2	69.3	69.9
	Merlin	43.1	50.3	54.9	56.3	57.9	59.6
	UMIVR	43.1	65.0	74.3	81.3	84.9	87.3
	Ours	43.1	**70.2**	**83.7**	**89.8**	**92.4**	**93.5**
BRI↓	IVR	-	1.050	0.876	0.793	0.742	0.707
	Merlin	-	1.230	1.127	1.054	0.999	0.953
	UMIVR	-	0.985	0.751	0.614	0.519	0.451
	Ours	-	**0.946**	**0.675**	**0.524**	**0.429**	**0.365**
AVSD-1k-Test Dataset							
R@1↑	IVR	30.3	36.0	38.8	40.6	41.3	41.5
	Merlin	30.3	34.9	37.5	38.9	40.3	41.3
	UMIVR	30.3	43.2	50.4	56.0	59.9	63.2
	Ours	30.3	**47.6**	**64.2**	**71.3**	**74.8**	**76.6**
BRI↓	IVR	-	1.802	1.698	1.638	1.599	1.570
	Merlin	-	1.838	1.743	1.670	1.613	1.566
	UMIVR	-	1.665	1.438	1.272	1.146	1.050
	Ours	-	**1.580**	**1.265**	**1.049**	**0.912**	**0.819**

Table 16: **Effect of backbone VLMs** on WebVid-CoVR-Test. ReCoVR shows consistent multi-turn improvement across two model families (Qwen3-VL and InternVL3.5) and two model sizes (4B and 8B). These results support the framework's backbone-agnostic design.

	• **BaseVLMs**	*Turn 1*	*2*	*3*	*4*	*5*
R@1↑	Qwen3-VL-4B	74.30	80.32	84.39	87.72	90.22
	Qwen3-VL-8B	74.84	81.69	85.80	89.12	91.88
	InternVL3.5-4B	75.59	82.94	87.6	90.61	92.64
	InternVL3.5-8B	77.74	84.31	88.30	91.00	93.00
R@5↑	Qwen3-VL-4B	87.28	89.28	91.28	93.04	94.13
	Qwen3-VL-8B	89.48	90.45	92.10	93.43	94.60
	InternVL3.5-4B	87.91	91.71	93.86	94.99	96.09
	InternVL3.5-8B	88.15	92.06	94.09	95.42	96.21
R@10↑	Qwen3-VL-4B	93.08	93.81	94.56	95.27	96.24
	Qwen3-VL-8B	93.43	94.01	95.11	96.40	96.91
	InternVL3.5-4B	91.55	95.19	96.32	96.87	97.57
	InternVL3.5-8B	91.74	95.15	96.40	97.26	97.77
BRI↓	Qwen3-VL-4B	0.6737	0.5271	0.4538	0.4034	0.3643
	Qwen3-VL-8B	0.6483	0.4884	0.4111	0.3601	0.3224
	InternVL3.5-4B	0.6365	0.4653	0.3803	0.3260	0.2868
	InternVL3.5-8B	0.6271	0.4543	0.3708	0.3172	0.2784

Generalization to interactive text-to-video retrieval. We evaluate whether the closed-loop Reflection Pathway transfers beyond ICoVR to standard interactive text-to-video retrieval (I-TVR). To isolate this effect, we integrate the Reflection Pathway into UMIVR as a plug-in module, while keeping the original I-TVR setting and feedback protocol. The interaction terminates once the target is found; otherwise, the simulated user follows the VideoQA-based feedback protocol used in prior I-TVR methods (Liang & Albanie, 2023; Han et al., 2024; Zhang et al., 2025).

As shown in Table 15, adding the Reflection Pathway consistently improves the open-loop UMIVR pipeline on both MSRVTT-1kA and AVSD-1k-Test. The gains are especially large on AVSD-1k-Test, where the lower Turn 0 accuracy leaves more room for trajectory-level correction. This suggests that the closed-loop diagnosis mechanism is not specific to composed retrieval, but can also strengthen existing interactive pipelines.

Generalization across reasoning VLMs. We evaluate whether ReCoVR depends on a particular reasoning VLM. In this experiment, the user simulator is fixed to Qwen3-VL-4B, while ReCoVR uses different reasoning VLMs from two families and two scales: Qwen3-VL-4B/8B and InternVL3.5-4B/8B. This setting separates the simulator from the retrieval agent and tests whether the observed gains require using the same VLM family as the simulator.

Table 16 shows that all reasoning VLMs produce steady multi-turn improvements. Larger models generally achieve higher final performance, but switching to a different model family also preserves the gain pattern. This suggests that ReCoVR's improvement comes from the closed-loop reasoning structure rather than simulator-agent model sharing. The results further indicate that the framework is robust to the choice and scale of the reasoning VLM. Qwen3-VL-4B is one of the smallest models in this comparison and is the default for every main-paper result. The reported gains therefore do not require the larger 8B variants.

C Efficiency Analysis

The closed loop comes at a computational price: every ReCoVR turn runs intent decomposition and trajectory reflection on the VLM, adding about 1.6s per executed turn over the strongest adapted baseline. Faster convergence partially offsets this overhead, but overall system time remains higher. Two properties keep this cost manageable. First, the overhead lies in reasoning rather than retrieval: nearest-neighbor search and rank fusion together take hundredths of a second per turn, so the cost scales with the number of VLM calls rather than with gallery size, and shrinks with a smaller or faster reasoning model. Second, reading results and writing feedback dominate each turn in the real-user latency estimate. Under a conservative estimate of human response time, the entire ReCoVR system accounts for roughly a seventh of end-to-end latency. The following analysis reports the per-turn breakdown, memory footprint, and deployment estimate.

All efficiency experiments in this section are conducted on a **single NVIDIA RTX A6000 GPU**. We evaluate on 100 instances sampled from WebVid-CoVR-Test (the same set used in the Sec. 4.4) under the same setup as the main experiments (up to five interactive turns). Among them, 57 instances fail at Turn 0 and thus enter the interactive stage. The interaction terminates early once the target video is ranked first, so fewer executed turns directly indicate faster convergence. Unless otherwise specified, all reported timings are measured from the interactive pipeline used in automatic evaluation. For ReCoVR, the full first-turn query reform cost is included in the reported Turn 1 latency, matching the definition used throughout the paper.

Table 17: **Method-level interactive cost on 100 instances.** We report total interactive time, the number of executed turns, the average number of executed turns per active query, and the average end-to-end time per executed turn.

Method	Total Time (s, 100q)	Executed Turns	Avg Turns / Active Query	Avg Time / Executed Turn (s)
IVR	482.06	223	3.91	2.16
UMIVR	637.15	231	4.05	2.76
ReCoVR (ours)	908.04	208	3.65	4.37

Table 17 provides a view of end-to-end efficiency beyond a simple average with 100 instances, since the number of executed turns varies under early stopping. ReCoVR has the largest total runtime, but executes the fewest interactive turns overall: 208 turns, compared with more than 220 turns for both adapted baselines. Thus, its larger total runtime is not due to longer interaction trajectories, but to higher per-turn computation. This is expected because each ReCoVR turn performs dual-channel retrieval and closed-loop reasoning, whereas the baselines use lighter open-loop query updates. Overall, ReCoVR trades higher per-turn cost for faster convergence in terms of executed turns.

Table 18 explains where the per-turn latency of ReCoVR comes from. In the automatic pipeline, the largest single component is the simulated user, which takes 1.77s per turn and accounts for 40.48% of the total latency. On the model side, the main cost comes from language-driven reasoning modules rather than retrieval itself: *Intent Decomposer* takes 0.95s, *Reflection* takes 0.85s, and *Query Reformer* takes 0.62s per turn, whereas the combined cost of T2V retrieval, CoVR retrieval, and TW-RRF is only about 0.03s. This confirms that the computational overhead of ReCoVR is dominated by its reasoning and diagnosis stages, not by nearest-neighbor retrieval.

More importantly, Table 18 shows why raw simulator timing should not be over-interpreted for realistic deployments. If the simulated user is replaced by a conservative estimate of 15 seconds of real human feedback per turn, the end-to-end turn cost becomes 17.60s. Under this setting, the user contributes about 85.24% of the latency budget, while the entire ReCoVR system contributes only 14.76%. ReCoVR is computationally heavier than the baselines. Human response time nevertheless dominates practical interaction time under this estimate.

Table 19 shows that the memory bank adds only a small footprint relative to model residency. For all methods, the bank primarily stores the long-term Progress Memory PM^L, which caches precomputed video captions for the gallery. ReCoVR's bank is slightly larger (3.50MB in RAM, 2.01MB on disk) because PM^L additionally stores richer metadata used by the Reflection Pathway, compared with the caption-only banks of IVR and UMIVR (0.98MB). In both cases, the bank is negligible relative to the $\sim$1.86GB base model footprint, occupying less than 0.19% of model memory. The short-term Progress Memory PM^S and the Result Memory RM are session-level structures that are created and discarded per query, contributing no persistent storage cost.

The table also reports the offline cost of bank construction on cache misses. ReCoVR requires 1.658s on average to build one PM^L entry, while IVR and UMIVR require 0.908s. Importantly, these are offline amortized costs: once built, the cached entries are reused across all sessions and queries, so they do not affect the online per-turn latency reported in Tables 17 and 18.

Overall, ReCoVR is not the fastest method in raw system time, which is the expected cost of dual-channel retrieval and closed-loop reasoning. However, it converges in fewer turns, its retrieval operations are extremely lightweight, and its bank footprint is negligible. Most importantly, in realistic use the dominant latency comes from human feedback rather than model computation, making ReCoVR's efficiency trade-off well justified by its retrieval gains.

Table 18: **Per-turn latency breakdown of ReCoVR.** Time is averaged over all executed interactive turns (N=208). Percentages are computed against the measured end-to-end turn cost in the automatic pipeline. Individual module percentages may not sum exactly to the subtotal due to rounding.

Module	Avg Time / Turn (s)	Ratio in Measured Pipeline (%)
Question Generation	0.158	3.62
Intent Decomposer	0.950	21.76
Query Reformer	0.619	14.17
T2V Retrieval	0.017	0.39
CoVR Retrieval	0.011	0.25
TW-RRF	0.001	0.02
Reflection	0.847	19.41
Simulated User Feedback	1.767	40.48
System Total (ReCoVR only)	2.598	59.52
Measured Total / Turn	4.365	100.00
Estimated Total / Turn with Real Human Feedback (15s)	17.598	-

Table 19: **Memory-bank footprint and offline bank-construction latency.** We report Resident Set Size (RSS) after model loading, after bank loading, and at peak during execution, along with the bank size in RAM and on disk, and the mean cost of constructing one bank entry on a cache miss.

Method	RSS After Model (MB)	RSS After Bank (MB)	RSS Peak (MB)	Bank RAM (MB)	Bank Disk (MB)	Build Mean (s)
IVR	1857.95	1940.98	2281.07	0.98	0.54	0.908
UMIVR	1857.91	1943.28	2274.54	0.98	0.54	0.908
ReCoVR (ours)	1857.81	1944.11	2157.97	3.50	2.01	1.658

D User Simulation and Human-Feedback Study Details

D.1 Motivation: Why Simulate Users?

Evaluating interactive retrieval systems faces a well-known tension between reproducibility and realism (Zhang et al., 2024; Dou et al., 2025). Offline evaluation on static benchmarks (e.g., measuring Recall or NDCG per turn) assumes that all preceding turns have been perfectly resolved, ignoring the cascading effects of errors across turns (Dalton et al., 2020). Online user studies capture realistic interaction dynamics but are expensive, difficult to scale, and hard to reproduce. This gap has motivated the development of *user simulators* that can interact with retrieval systems in a controlled, automated loop (Zhang et al., 2024; Azzopardi et al., 2007; Degachi et al., 2025; Aliannejadi et al., 2024).

In the information retrieval community, user simulation has evolved from early rule-based or probabilistic models that estimate click and query reformulation behaviors (Mo et al., 2025) from search logs (Azzopardi et al., 2007), to recent LLM-based approaches that leverage instruction following and reasoning capabilities to role-play as users with specific information needs (Zhang et al., 2024; Wang et al., 2024; Erbacher et al., 2022; Balog & Zhai, 2023). For instance, USimAgent (Zhang et al., 2024) employs an LLM to generate queries, predict clicks, and decide when to stop searching, demonstrating that LLM-based simulators can produce coherent multi-round search sessions that approximate real user behavior. These advances have made simulated-user evaluation a standard paradigm for interactive systems where large-scale human studies are impractical.

In interactive text-to-video retrieval (I-TVR), existing methods adopt a question-answering interaction protocol: the system generates a clarifying question at each turn, and a user simulator provides an answer based on the ground-truth target (Liang & Albanie, 2023; Han et al., 2024; Zhang et al., 2025). This protocol has enabled reproducible multi-turn evaluation across I-TVR benchmarks.

However, **composed video retrieval presents a fundamentally different challenge**. Unlike text-only retrieval where user intent is expressed through text alone, ICoVR requires the user to coordinate feedback across two modalities, referencing both visual content and textual modifications simultaneously. The user must observe a video result, compare it against a mental model of the target, and articulate the remaining visual and semantic gap, which demands a vision-language understanding capability that text-only simulators cannot provide. Furthermore, user feedback in ICoVR is not limited to answering predefined questions; it may take the form of relative modifi-

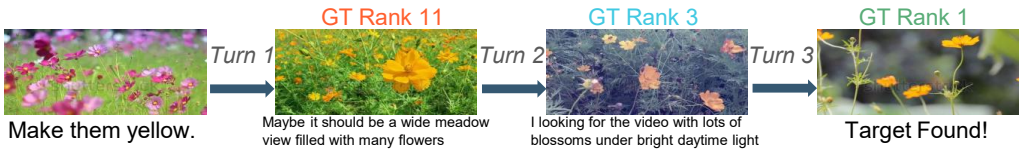

Figure 5: **Qualitative trajectory with human feedback.** A real user progressively refines the search intent across three turns, improving the ground-truth rank from 11 to 3 and finally to 1.

cations (e.g., "`make the scene brighter`"), direct descriptions (e.g., "`I want a beach at sunset`"), or answers to system-generated questions, requiring a flexible simulation framework. No prior work has addressed user simulation for the composed video retrieval setting. These considerations motivate us to design a user simulator tailored for ICoVR, which we describe next.

D.2 User Simulator Design

Our simulator is built on a vision-language model that acts as a surrogate user throughout the interactive session. The overall procedure is summarized in Algorithm 1. At the start of each evaluation query, the simulator receives the ground-truth target video v_{gt} and generates an internal caption to establish a persistent visual memory of the target (Line 1). The simulator caches this caption and the target video representation for reuse across all subsequent turns. This keeps its representation of the search goal consistent.

At each interaction turn t, the simulator receives the system's top-ranked candidate $\hat{L}_t(1)$ and optionally a clarifying question q_t. If the candidate matches the ground truth, the session terminates (Line 4). Otherwise, the simulator enters a two-stage process. In the *analysis stage* (Line 7), the VLM jointly observes both the target and the candidate video and produces a structured comparison, classifying the discrepancy as either `MINOR` (the candidate is semantically related but differs in specific attributes) or `MAJOR` (the candidate is largely irrelevant to the target). In the *feedback generation stage* (Lines 8–13), the simulator selects an action type based on the analysis outcome and generates natural-language feedback u_t accordingly. This follows recent user simulation protocols that separate target understanding from feedback generation (Zhang et al., 2024), while adapting the action space to the relative and descriptive feedback that composed retrieval invites. The action depends on the target, the displayed result, and whether a question was asked, not on the method being evaluated:

- **Modify**: When the discrepancy is minor, the simulator suggests a targeted modification to the current result (e.g., "`change the background to outdoor`").
- **Rewrite**: When the discrepancy is major, the simulator generates a fresh description of the target video.
- **Answer**: When a system question q_t is available and the discrepancy is major, the simulator may choose to answer the question based on its knowledge of the target.

This action selection produces a mix of relative and descriptive feedback, mirroring the diverse interaction patterns of real users. All methods in our experiments interact with the same simulator under identical prompting configurations, ensuring a fair comparison.

D.3 Human-Feedback Interface

To support the human-feedback evaluation of Sec. 4.4 of the main paper, we develop a web-based interaction interface (Fig. 6). The human-feedback study received institutional ethics approval. Participants were informed of the task procedure and minimal risks, and no demographic attributes or sensitive personal information were collected. The interface keeps the interaction setting close to the simulator-based evaluation protocol while allowing participants to provide free-form natural language feedback. At each turn, the participant views: (1) the reference video from the original CoVR instance, (2) the current top-ranked candidate returned by ReCoVR, (3) the initial modification text, and (4) an optional clarifying question generated by the system. A text box allows the participant to write natural-language feedback describing how the current result should be modified.

Algorithm 1 User Simulator for ICoVR

Require: Ground-truth target video v_{gt}; VLM $\mathcal{V}$; max turns T

caption $\leftarrow \mathcal{V}.\text{describe}(v_{\text{gt}})$ {Cache target representation}
for $t = 1$ **to** T **do**
  Receive candidate $\hat{L}_t(1)$ and optional question q_t from system
  if $\hat{L}_t(1) = v_{\text{gt}}$ **then**
    return `STOP`
  end if
  $(\text{status},\ \text{analysis}) \leftarrow \mathcal{V}.\text{compare}(v_{\text{gt}},\ \hat{L}_t(1))$ {status $\in$ {`MINOR`, `MAJOR`}}
  if status $=$ `MINOR` **then**
    $u_t \leftarrow \mathcal{V}.\text{modify}(\text{analysis},\ v_{\text{gt}})$
  else if $q_t \neq \varnothing$ **and** $\text{rand}() < 0.5$ **then**
    $u_t \leftarrow \mathcal{V}.\text{answer}(q_t,\ v_{\text{gt}})$
  else
    $u_t \leftarrow \mathcal{V}.\text{rewrite}(v_{\text{gt}})$
  end if
  Send feedback u_t to the retrieval system
end for

Table 20: **Robustness to degraded simulated feedback** on WebVid-CoVR-Test. We report Turn 1 R@1, Turn 5 R@1, the interactive gain from Turn 0 to Turn 5, and BRI at Turn 5. Among the genuinely degraded settings (*surface noise*, *constraint dropping*, and *modify-only*), ReCoVR remains consistently stronger than UMIVR and preserves large multi-turn gains. The *rewrite-only* setting is included as a strategy-shift control.

Category	Setting	UMIVR				ReCoVR (ours)			
		Turn 1	*Turn 5*	*Gain@5*	*BRI@5*	*Turn 1*	*Turn 5*	*Gain@5*	*BRI@5*
Clean	Default simulator	67.10	75.04	20.74	0.4700	74.30	90.22	35.92	0.3643
Surface Noise	Light: word swap + synonym replacement	66.98	74.65	20.35	0.4745	73.90	89.01	34.71	0.3757
	Heavy: word deletion + character swap	64.91	74.06	19.76	0.4978	71.40	88.73	34.43	0.3942
Constraint Dropping	Mild drop (p=0.3)	66.71	74.77	20.47	0.4818	72.50	89.28	34.98	0.3781
	Severe drop (p=0.5)	66.35	73.83	19.53	0.4876	72.30	89.08	34.78	0.3805
Behavioral Policy Shift	Modify-only	60.02	64.05	9.75	0.6726	69.48	86.15	31.85	0.4405
	Rewrite-only (control)	71.40	78.21	23.91	0.4023	76.72	92.14	37.84	0.2975

The participant can also click a "`Found It`" button to indicate that the target has been located, which terminates the session early.

An interaction history panel records the previous turns, including the system question, the participant's feedback, and the resulting top-ranked candidate. This allows the participant to track how the search evolves over the session. For this controlled setting, the ground-truth target video is displayed in a separate section at the bottom as the intended search goal. Each participant is assigned a unique user ID and completes up to 20 instances, with a maximum of 3 interaction turns per instance. Fig. 5 further illustrates a real interaction where the user progressively refines the search intent and guides ReCoVR to the target in three turns.

D.4 Robustness to Degraded Simulated Feedback

The main experiments use a target-conditioned simulator to provide scalable and repeatable feedback during offline evaluation. While this controlled protocol enables fair comparison across methods, real users may provide noisy, incomplete, sparse, or strategy-dependent feedback. We therefore stress-test ReCoVR by systematically degrading the simulated feedback on WebVid-CoVR-Test along three axes: *surface noise*, which corrupts wording via character- and word-level perturbations (Qiu et al., 2022); *constraint dropping*, which randomly removes partial semantic constraints from the feedback; and *behavioral policy shifts*, which replace the default mixed strategy with a single fixed interaction mode. These settings cover common non-ideal feedback conditions such as noisy wording, missing constraints, and sparse local edits.

Across all genuinely degraded conditions, ReCoVR's interactive gain remains within the narrow band of 31.85–34.98 points, compared with only 9.75–20.47 for UMIVR. Under surface-level perturbations and partial constraint omission, ReCoVR's Turn 5 R@1 stays within 1.49 points of its clean performance, with BRI degradation contained to at most 0.03. This tolerance is consistent

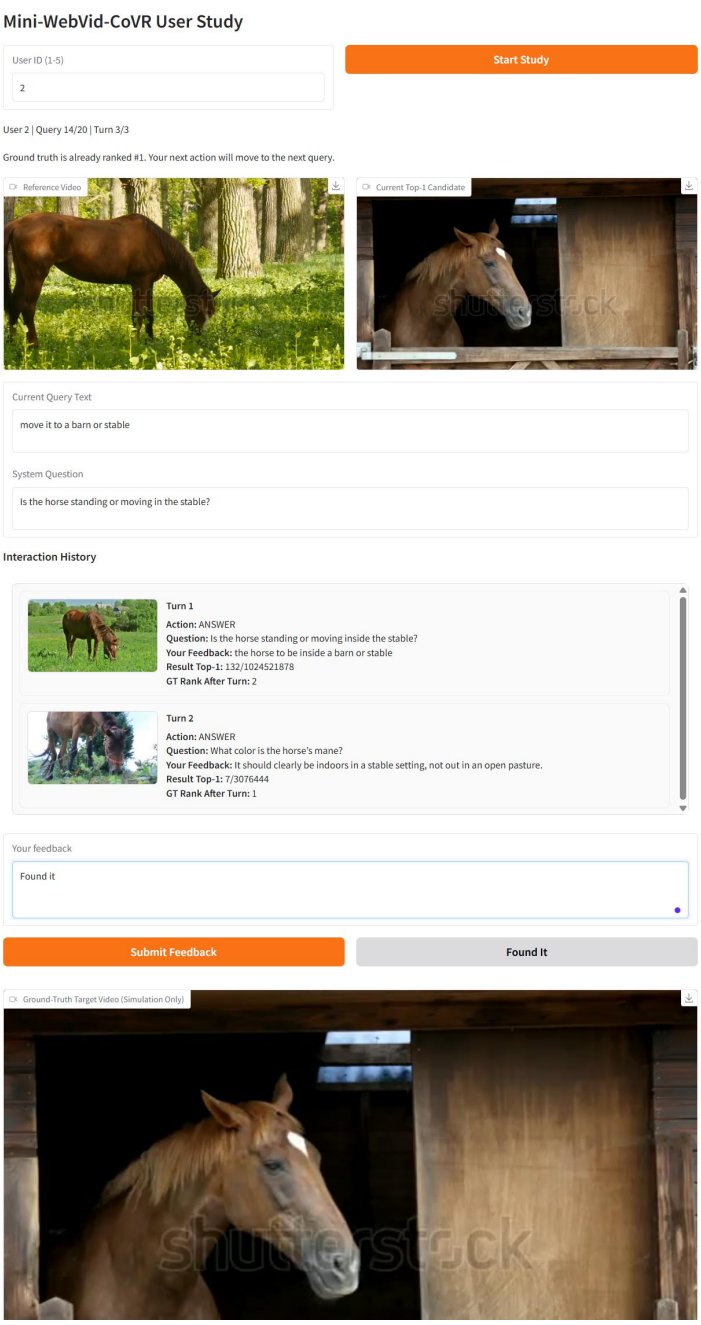

Figure 6: **Human-feedback interface.** Participants view the reference video, current top-ranked candidate, initial modification text, and an optional system question. They provide free-form feedback in the text box or click "Found It" to terminate the session. The interaction history panel tracks the search progression, and the ground-truth target video is shown at the bottom as the intended search goal.

with the design of the Reflection Pathway, which accumulates constraints across turns and corrects rankings from the retrieval trajectory rather than relying solely on the literal content of the latest feedback.

The most revealing test is the *modify-only* policy, which forces the simulator to provide only local edits while suppressing any direct description of the target video. This severely limits the informa-

tion available per turn. Under this condition, UMIVR's interactive gain collapses to just 9.75 points, whereas ReCoVR still achieves a gain of 31.85 points, reaching 86.15% R@1 at Turn 5. This gap directly reflects the value of the dual-channel Intent Pathway: even when user feedback contains only relative modifications, the T2V channel can still construct effective standalone queries from the accumulated semantic context in Progress Memory, compensating for the absence of explicit target descriptions.

By contrast, the *rewrite-only* policy improves both methods beyond their clean performance. Initial modification texts on WebVid-CoVR average only 4.6 words. A complete target description at each turn gives the retriever strictly more information than a partial edit. We include this condition as a strategy-shift control rather than a degradation case. The default simulator uses a mixed strategy to model the diversity of real user behavior rather than to maximize retrieval performance under any single mode.

Both behavioral policy conditions also bound how much ReCoVR depends on its optional Question Asking module. In these conditions the simulator follows a fixed interaction mode and never conditions on the generated question q_t, so they are equivalent to running the system with question asking disabled. ReCoVR reaches 86.15% and 92.14% R@1 at Turn 5 in the two cases, which indicates that the closed loop is driven by the retrieval trajectory rather than by whether the user chooses to answer.

These stress tests complement the human-feedback evaluation in Table 4 of the main paper. Human participants provide free-form and less template-like feedback, while the degraded simulator tests systematically noisy, incomplete, and sparse feedback conditions. Together, they indicate that ReCoVR's advantage does not depend on perfectly informative feedback.

E Qualitative Case Studies

We present two successful examples and one same-instance baseline comparison to illustrate ReCoVR's behavior in practice. Fig. 7 shows a WebVid-CoVR query where the user starts with "`make it sunrise`" (GT rank 42) and progressively specifies the scene across two turns. The Reflection Pathway first extracts a negative constraint to suppress irrelevant cloud types (σ_1=neutral), then accumulates positive constraints [*airplane wing*,...] after an explicit rejection (σ_2=negative), driving TW-RRF to rank 1. In contrast, Fig. 8 shows the same target instance under UMIVR, which remains stuck at rank 5 despite explicit scene, airplane-wing, and viewpoint feedback. Fig. 9 illustrates a FineCVR case with a complex initial modification text (GT rank 7). Here the CoVR channel plays the dominant role: fine-grained appearance and action cues ("`red cap`," "`door opening`") are best expressed relative to the visual reference, allowing the CoVR channel to reach rank 1 at Turn 2. Together, these cases demonstrate the complementarity of the dual channels and the role of reflection: T2V excels with holistic scene descriptions, CoVR is more effective for fine-grained reference-anchored refinements, and trajectory-level reflection helps correct stalled retrieval.

Fig. 10 presents a case where the reference video (indoor, male subject) differs drastically from the target (outdoor, blonde female near a green sign). Despite multiple turns of user feedback, the CoVR channel fails to retrieve the target, ending at rank 168 because composed retrieval anchors to a visual reference that shares almost no similarity with the target. Although the T2V channel reaches rank 1 by the final turn, the fixed fusion strategy cannot suppress the CoVR channel's noisy contribution, yielding a final TW-RRF rank of only 22. This exposes a limitation of the current design: when the reference-target gap is large, adaptive channel weighting that down-weights the less reliable channel could improve fusion quality, which we leave for future work.

F Algorithmic Overview of ReCoVR

The method description in the main paper is organized by pathway. Algorithm 2 summarizes the end-to-end execution flow. It shows the data dependencies among modules and when each memory component is read and updated.

At each interactive turn, the Intent Pathway (Lines 14–16) and the Reflection Pathway (Lines 17–20) read the same snapshot PM_{t-1}^{S} and can therefore execute independently. Their outputs converge at the retrieval stage: Intent determines *what to search for* via the dual channels (Lines 21–23), while

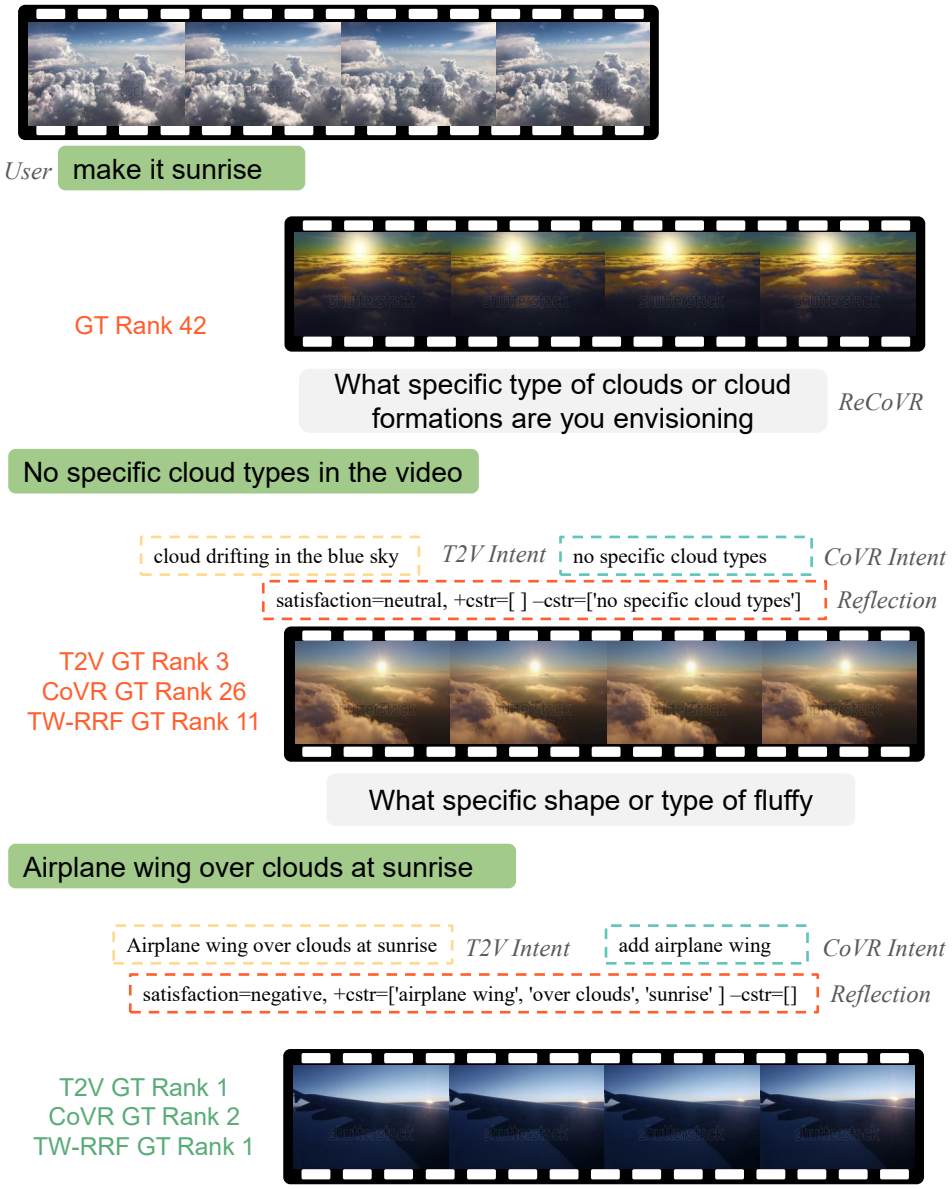

Figure 7: **Successful case on WebVid-CoVR.** The user progressively specifies scene details across two turns, improving the target from rank 42 to rank 1. The T2V channel drives convergence after the user provides a holistic scene description in Turn 2.

Reflection determines *what to suppress* via the rank-cap (Line 24). TW-RRF (Line 26) then fuses the full trajectory stored in RM into the final ranking, after which both memory components are updated to prepare the state for the next turn.

G Prompt Templates

To improve reproducibility, we provide the prompt templates used by the main reasoning modules of ReCoVR, together with the prompt template used by the simulated user. Tables 21–24 correspond to the four reasoning modules of ReCoVR (Intent Decomposer, Query Reformer, State Reflector, and Question Asking, respectively), following the execution order described in Algorithm 2. Table 25

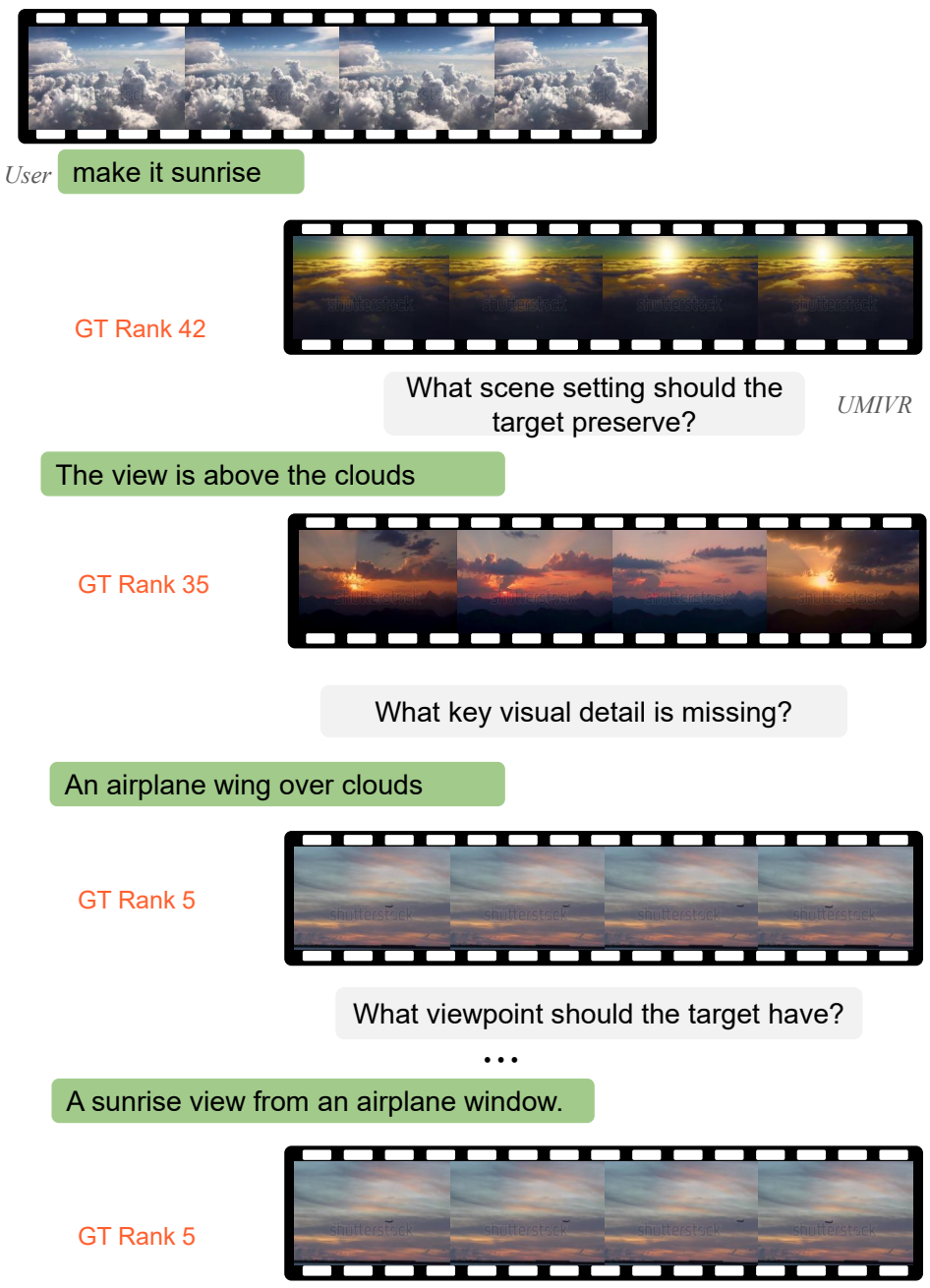

Figure 8: **Comparison on the same WebVid-CoVR case.** While ReCoVR improves the target from rank 42 to rank 1 in Fig. 7, UMIVR remains stuck at rank 5 under the same target instance and interaction budget, despite receiving explicit scene, airplane-wing, and viewpoint feedback. This comparison highlights the importance of trajectory-level reflection for correcting stalled retrieval.

lists the prompts used by the simulated user (Sec. D.2); all methods in our experiments share this same simulator configuration.

Two design principles run through all templates. First, each prompt enforces a strict *role boundary*: modules are instructed to decompose, compress, or analyze without hallucinating details absent from the input, preventing error amplification across turns. Second, the Reflection prompt explicitly requests *incremental* constraints, avoiding redundancy with the accumulated session state in PM^S.



Figure 9: **Successful case on FineCVR.** Fine-grained appearance and action cues are anchored to the visual reference via the CoVR channel, improving the target from rank 7 to rank 1 over two interaction turns.

H Limitations and Discussion

Evaluation Scope and Interaction Variability. This work focuses on the algorithmic side of interactive retrieval: how feedback, retrieval history, and intermediate results should be used to update rankings across turns. Modeling the full diversity of real user behavior is a broader HCI-oriented research direction (Balog & Zhai, 2023; Kelly, 2009; Marchionini, 2006; Preece et al., 1994). It covers search strategies, patience, expertise, and interface effects. We therefore use a controlled interactive setting to isolate algorithmic differences under shared targets, initial rankings, feedback sources, and stopping criteria, and further complement it with a free-form human-feedback evaluation and degraded-feedback analysis. Larger-scale user studies with diverse behavioral profiles (Balog & Zhai, 2023; Zhang et al., 2024; Mo et al., 2025) remain an important direction for future work. More broadly, ICoVR inherits the open-ended nature of composed retrieval, where interaction progressively reduces ambiguity among plausible targets rather than eliminating it entirely.

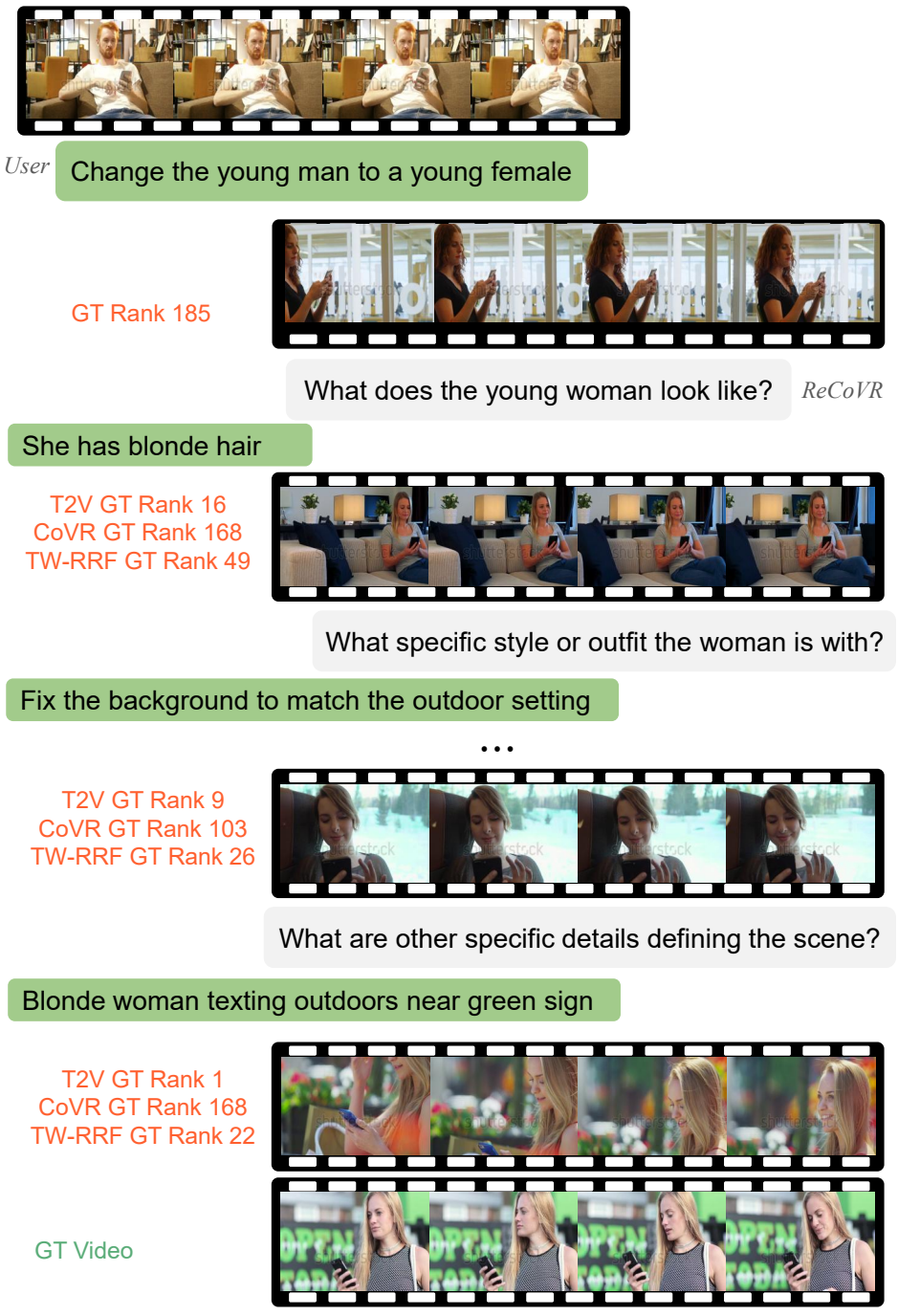

Figure 10: **Failure case.** A large visual gap between the reference (indoor, male) and target (outdoor, female) leaves the CoVR channel with a final target rank of 168. Despite the T2V channel reaching rank 1, fixed fusion yields a final rank of only 22.

Design Trade-offs and Deployment. ReCoVR uses simple, training-free, rank-level operations to remain backbone-agnostic and easy to plug into existing retrievers, and our ablations and sensitivity analyses suggest that the gains are not tied to narrow hyperparameter choices. Future work could make fusion more adaptive by estimating per-turn channel reliability from cross-channel agreement, query–reference consistency, or uncertainty signals. ReCoVR also introduces additional per-turn reasoning cost for intent decomposition and trajectory reflection; although this is partly offset by

Algorithm 2 ReCoVR: Interactive Composed Video Retrieval

Require: Reference video v_{ref}, modification text u_0, gallery $\mathcal{G}$, max turns T
Ensure: Final ranked list $\hat{L}$

{*Initialization*}
$PM^L \leftarrow \text{CacheGalleryCaptions}(\mathcal{G})$
$PM^S \leftarrow \{v_{\text{ref}},\ u_0,\ +\text{cstr}{=}\varnothing,\ -\text{cstr}{=}\varnothing\}$
$RM \leftarrow \varnothing$
{*Turn 0: standard single-turn CoVR*}
$\hat{L}_0 \leftarrow \text{CoVR}(v_{\text{ref}},\ u_0,\ \mathcal{G})$; $RM \leftarrow RM \cup \{(0,\ \hat{L}_0)\}$
for $t = 1$ **to** T **do**
  {*User interaction*}
  $q_t \leftarrow \text{AskQuestion}(PM^S)$
  Present $\hat{L}_{t-1}(1)$ and q_t to the user; receive feedback u_t
  if the user accepts $\hat{L}_{t-1}(1)$ **then**
    return $\hat{L}_{t-1}$
  end if
  {*Intent Pathway (uses PM^S_{t-1} which includes Δc_{t-1})*}
  $(\text{search}_t,\ \text{edit}_t) \leftarrow \text{IntentDecomposer}(u_t,\ PM^S,\ PM^L)$
  $(\text{search}_t,\ \text{edit}_t) \leftarrow \text{QueryReformer}(\text{search}_t,\ \text{edit}_t,\ PM^S)$
  {*Reflection Pathway (independent of current Intent)*}
  $(\sigma_t,\ \Delta c_t) \leftarrow \text{StateReflector}(u_t,\ PM^S,\ PM^L)$
  $\Delta r_t \leftarrow \text{RankCapPolicy}(\sigma_t,\ RM)$ {demote rejected/stagnant candidates}
  $PM^S \leftarrow \text{UpdateConstraints}(PM^S,\ \Delta c_t)$
  {*Dual-channel retrieval*}
  $L_t^{\text{T2V}} \leftarrow \text{T2V}(\text{search}_t,\ \mathcal{G})$
  $L_t^{\text{CoVR}} \leftarrow \text{CoVR}\big(\hat{L}_{t-1}(1),\ \text{edit}_t,\ \mathcal{G}\big)$ {dynamic anchoring}
  $(L_t^{\text{T2V}},\ L_t^{\text{CoVR}}) \leftarrow \text{ApplyRankCap}(L_t^{\text{T2V}},\ L_t^{\text{CoVR}},\ \Delta r_t)$
  {*Fusion and memory update*}
  $RM \leftarrow RM \cup \{(t,\ L_t^{\text{T2V}},\ L_t^{\text{CoVR}})\}$
  $\hat{L}_t \leftarrow \text{TW-RRF}(RM)$
  $PM^S \leftarrow \text{UpdateProgress}(PM^S,\ \text{search}_t,\ \text{edit}_t,\ \hat{L}_t(1))$
end for
return $\hat{L}_T$

fewer executed turns, real-time deployment could benefit from caching, smaller VLMs, distillation, or lightweight structured parsers.

Scope of the Reported Evidence. We summarize how each analysis should be interpreted.The paired intervals of Table 7 quantify test-session sampling variability under a fixed model, simulator, and protocol; they are not tests over independent turns, and we report no p-values. The failure diagnostics of Sec. A.4 are temporal associations measured on executed trajectories: they show that ReCoVR recovers where the open-loop baseline stalls, but a single run cannot attribute that recovery to one module in isolation, which is why component attribution rests on the matched ablations instead. The rank-worsening row is an indicator based on target rank, not a full semantic measurement of query drift. The structured-output statistics describe interface reliability rather than the semantic correctness of each decision. Natural extensions include reasoning models below the 4B scale, user studies with broader participant pools, and latency measurements with question generation disabled. We do not report an oracle anchor, since it would require target-side information that is unavailable at retrieval time.

Relation to LLM Agent Frameworks. ReCoVR shares structural similarities with agent frameworks such as ReAct (Yao et al., 2022) and Reflexion (Shinn et al., 2023), but is purpose-built for interactive retrieval. Its action space is constrained to query reformulation, dual-channel retrieval, rank-level intervention, and rank fusion, while its reflection is grounded in ranked-list trajectories

Table 21: **Prompt template for the Intent Decomposer** (Sec. 3.2 of the main paper). Given user feedback u_t, the module decomposes it into a target-descriptive search query and an optional edit instruction, corresponding to the T2V and CoVR channels respectively.

Field	Content
System prompt	`null`

Template	`Task: Decompose user feedback into search and edit components for interactive video retrieval.`
	`Context:`
	`- Previous question asked to user: {prev_question}`
	`- Previous top-1 retrieved video description: {prev_top1_caption}`
	`- User feedback: {user_query}`
	`Role boundary:`
	`- Only decompose intent.`
	`- Do NOT rewrite, enrich, or add details not present in feedback.`
	`Steps:`
	`1. Determine whether the feedback mainly answers the previous question.`
	`2. Determine whether the feedback contains explicit edit intent (change/remove/replace/add/make/swap/turn/switch/modify, etc.).`
	`3. Extract search_query_info as target-descriptive information (objects/actions/scenes) when available.`
	`4. Extract edit_query_info only when edit intent exists.`
	`5. If feedback is purely an answer and no clear standalone target description is provided, search_query_info can be empty.`
	`6. If no edit intent, set edit_query_info to an empty string.`
	`Return JSON only: {output_schema}`

Output schema	`{"reasoning": "string (1-2 sentences)", "is_answer_to_prev_question": "boolean", "has_edit_intent": "boolean", "search_query_info": "string", "edit_query_info": "string"}`

rather than free-form critique. Reasoning traces may help single-turn edit understanding, but do not directly specify how to monitor drift, suppress repeated failures, or fuse evolving evidence across turns.

Broader Impacts. This work may improve interactive access to large-scale video collections in applications such as media search, education, and assistive information retrieval. However, stronger video retrieval systems should be deployed carefully when applied to sensitive or personal video collections, where unauthorized search or unintended exposure may raise privacy concerns. Responsible deployment should consider data consent, access control, and dataset/model biases.

Table 22: **Prompt templates for the Query Reformer** (Sec. 3.2 of the main paper). Two submodules serve different purposes: `t2v_initialize` generates the initial T2V search query at Turn 0 by applying u_0 to the reference video, while `rewrite_queries` compresses the accumulated running query when it exceeds the length threshold at Turn $t \geq 1$.

Field	Content
Module	`t2v_initialize` (Turn 0 only)
System prompt	`null`

Template	`Task: Generate a text-to-video search query from a reference video and an edit instruction.`
	`You see a reference video. The user wants the target video after applying`
	`the edit instruction.`
	`Edit instruction: {edit_query}`
	`Requirements:`
	`1. Use only visible video content plus the given edit instruction.`
	`2. Produce one concise target description for retrieval.`
	`3. Keep key visual details (objects, actions, scene, attributes)`
	`while avoiding unnecessary verbosity.`
	`Return JSON only: {output_schema}`

Output schema	`{"reasoning": "string (1 sentence)", "search_query": "string"}`
Module	`rewrite_queries` (Turn $t \geq 1$, triggered when query exceeds length threshold)
System prompt	`null`

Template	`Task: Compress a long retrieval query while preserving all critical visual information.`
	`Query to compress: {query_to_rewrite}`
	`Requirements:`
	`1. Preserve key objects, actions, scenes, attributes, and constraints.`
	`2. Remove redundancy and keep retrieval semantics intact.`
	`3. Return one concise rewritten query.`
	`Return JSON only: {output_schema}`

Output schema	`{"reasoning": "string (1 sentence)", "rewritten_query": "string"}`

Table 23: **Prompt template for the State Reflector** (Sec. 3.2 of the main paper). The `reflection_context` variable is populated from PM^S_{t-1} (accumulated queries, prior positive/negative constraints) and relevant gallery captions from PM^L, corresponding to the inputs (u_t, PM^S_{t-1}, PM^L) in Eq. (4).

Field	Content
System prompt	`null`

Template	`Task: Analyze user feedback on retrieval results to assess satisfaction`
	`and extract incremental constraints.`
	`Input feedback: {user_feedback}`
	`Context: {reflection_context}`
	`Instructions:`
	`1. Decide satisfaction_level in {negative, neutral}.`
	`2. Use "negative" only for explicit rejection (e.g., wrong/not this/not desired).`
	`3. Extract NEW positive constraints (what user wants) as short phrases.`
	`4. Extract NEW negative constraints (what user wants to avoid) as short phrases.`
	`5. Keep outputs incremental: do not repeat constraints already implied by context`
	`unless the user explicitly strengthens them.`
	`6. Do not hallucinate unstated details.`
	`Return JSON only: {output_schema}`

Output schema	`{"reasoning": "string (1-2 sentences)",`
	`"satisfaction_level": "negative | neutral",`
	`"positive_constraints": ["string"],`
	`"negative_constraints": ["string"]}`

Table 24: **Prompt template for Question Asking** (Sec. 3.2 of the main paper). The module identifies remaining ambiguities from the current session state and generates one clarifying question per turn.

Field	Content
System prompt	`You generate one concise clarifying question for interactive video retrieval.`

Template	`Current search query: "{current_search_query}"`
	`Current edit query: "{current_edit_query}"`
	`Current constraints:`
	`- positive: {constraints_positive}`
	`- negative: {constraints_negative}`
	`Ask one short open-ended question that helps disambiguate remaining`
	`uncertainty in target appearance, actions, or scene.`
	`Return ONLY the question.`

Output schema	`{"reasoning_brief": "string (1 sentence: what ambiguity you identified)",` `"question": "string (one short question)"}`

Table 25: **Prompt templates for the simulated user** (Sec. D.2). The simulator follows a two-stage protocol: Stage 1 classifies the discrepancy between the target and the candidate (`MINOR_DIFF` and `MAJOR_DIFF` correspond to the `MINOR` and `MAJOR` status in Algorithm 1, respectively); Stage 2 selects one of three feedback policies accordingly. All methods in our experiments share this identical simulator.

Field	Content
System persona	`You are a casual user searching for a video. You give short, direct feedback.` `Do not explain your reasoning. Keep responses under 10 words.`
Target caption prompt	`Describe this target video in one short sentence. This is the video you should remember.`

Stage 1: analysis	`Compare two DIFFERENT videos step by step.`
	`Video 1 = TARGET (what I want)`
	`Video 2 = CANDIDATE (what was found)`
	`Step 1: What is the KEY DIFFERENCE between the two videos?`
	`Step 2: Do they share the SAME CATEGORY? (same animal type, same object type, same scene type)`
	`Step 3: Decide status based on Step 2:`
	`- If same category (e.g. goat vs goat, nuts vs nuts, river vs river) → MINOR_DIFF`
	`- If different categories (e.g. dog vs car) → MAJOR_DIFF`
	`Output JSON with fields in this order:`
	`{"diff_analysis": "...", "same_category": true/false, "status": "MINOR_DIFF" or "MAJOR_DIFF"}`

Stage 2: modify	`Video 2 is close but has specific errors (e.g., wrong color, wrong object).`
	`Give a short command to fix it. Example: "Change the red car to a blue one"`
	`or "Make it sunny".`

Stage 2: rewrite	`The candidate video is completely unrelated to the target. Ignore the candidate.`
	`Write a new, stand-alone search query that describes Video 1 (Target).`
	`Be descriptive but natural. Example: "A dog running on the beach under the sun".`

Stage 2: answer	`The search system asked: "{question}". Answer based on Video 1 (Target).`
	`Answer directly and briefly. Example: "It is red", "He is running".`
	`Do not speak in full paragraphs.`